SPRINGER BRIEFS IN LINGUISTICS

Monika S. Schmid · Sanne M. Berends
Christopher Bergmann · Susanne M. Brouwer
Nienke Meulman · Bregtje J. Seton
Simone A. Sprenger · Laurie A. Stowe

Designing Research on Bilingual Development

Behavioral and Neurolinguistic Experiments

EXTRAS ONLINE

Springer

SpringerBriefs in Linguistics

Series editor

Helen Aristar-Dry, Linguist List, Ypsilanti, MI, USA
and Dripping Springs, TX, USA

More information about this series at http://www.springer.com/series/11940

Monika S. Schmid · Sanne M. Berends
Christopher Bergmann · Susanne M. Brouwer
Nienke Meulman · Bregtje J. Seton
Simone A. Sprenger · Laurie A. Stowe

Designing Research on Bilingual Development

Behavioral and Neurolinguistic Experiments

Springer

Monika S. Schmid
Centre for Research in Language Development throughout the Lifespan (LaDeLi), Department of Language and Linguistics
University of Essex
Colchester
UK

Sanne M. Berends
Center for Language and Cognition, Faculty of Arts
University of Groningen
Groningen
The Netherlands

Christopher Bergmann
Center for Language and Cognition, Faculty of Arts
University of Groningen
Groningen
The Netherlands

Susanne M. Brouwer
Department of Special Education: Cognitive and Motor Disabilities
University of Utrecht
Utrecht
The Netherlands

Nienke Meulman
Center for Language and Cognition, Faculty of Arts
University of Groningen
Groningen
The Netherlands

Bregtje J. Seton
Center for Language and Cognition, Faculty of Arts
University of Groningen
Groningen
The Netherlands

Simone A. Sprenger
Center for Language and Cognition, Faculty of Arts
University of Groningen
Groningen
The Netherlands

Laurie A. Stowe
Center for Language and Cognition, Faculty of Arts
University of Groningen
Groningen
The Netherlands

Additional material to this book can be downloaded from http://extras.springer.com.

ISSN 2197-0009 ISSN 2197-0017 (electronic)
SpringerBriefs in Linguistics
ISBN 978-3-319-11528-3 ISBN 978-3-319-11529-0 (eBook)
DOI 10.1007/978-3-319-11529-0

Library of Congress Control Number: 2015943033

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Contents

Chapter 1
Introduction

Monika S. Schmid

Abstract The present text addresses theoretical and practical concerns that are relevant for large-scale investigations of bilingual development. It discusses the necessity of approaches that use a variety of elicitation methods and assess different populations. Such investigations can help resolve some of the most important current questions and controversies in the field of bilingualism, but they come with a number of practical and methodological challenges. We will set out some of these issues as we have encountered them in an investigation of bilingual development which was conducted over two continents and four countries, comprised a variety of elicitation techniques, from neuroimaging experiments to behavioral tasks and investigated populations from varied language backgrounds. This documentation of our approach is intended to help researchers who face similar challenges.

Keywords Multi-lab investigations of bilingualism · Second language acquisition · Language attrition · Multi-method studies

1.1 Types of Bilingualism

The nature of bilingualism and bilingual knowledge is one of the most elusive phenomena in present-day research on the human mind and human cognition. Different views on bilingual development abound and are highly controversial, ranging from the position that learning a first language (L1) from birth and a second language (L2) later in life are 'fundamentally different' (Bley-Vroman 2009) to approaches claiming that the processes, learning mechanisms, processing routines and other cognitive underpinnings are in principle the same (MacWhinney 2012).

Electronic supplementary material The online version of this chapter (doi:10.1007/978-3-319-11529-0_1) contains supplementary material, which is available to authorized users.

M.S. Schmid et al., *Designing Research on Bilingual Development*,
SpringerBriefs in Linguistics, DOI 10.1007/978-3-319-11529-0_1

One of the reasons for this multitude of views is the fact that types of bilingualism are as varied as the individual bilingual speaker. This multiplicity is evident in the fact that even linguists working in very closely related areas do not agree on how to use certain extremely common labels. For some, the term ‘bilingualism’ has an extremely narrow meaning, encompassing only those individuals who were exposed more or less equally to two (or more) languages from birth or early childhood and/or who speak both at the same level of proficiency. In other investigations, however, ‘bilingualism’ is a term that includes everyone who is able to communicate at a basic level in more than one language, i.e., to use more than one grammar creatively (Paradis 2009: 2). Researchers may also differentiate between terms such as ‘learning’ and ‘acquisition’, reserving the former for explicit, conscious (i.e., instructed) contexts and the latter for implicit (naturalistic) settings (Paradis 2009: 11) while others use them interchangeably. These inconsistencies may seem comparatively trivial issues of labeling, but they signal the fact that bilingual development is dependent on a host of factors, including but by no means limited to age and context of learning, which may have a dramatic impact on both the developmental trajectory and its outcome. Presumably those researchers who favor differentiated terms as characterized by different levels of these factors (simultaneous vs. sequential learning, instructed vs. naturalistic learning) usually also ascribe a qualitative impact to them, while those who opt for overarching labels consider them merely quantitatively distinct. In this text, we adopt a broad use of the term ‘bilingual’ to encompass anyone who has the ability to communicate in more than one language. We do not restrict the term in any way to a specific age of onset, learning context, proficiency level, or number of languages known (i.e., we also subsume multilinguals under this label). We also do not distinguish between the terms ‘learning’ and ‘acquisition’, as we make no a priori claims about what each of these factors contributes to a complete theory of bilingualism.

A substantial problem for linguistic research is posed by the fact that the predictors which have been claimed by some researchers to play a major role in determining both path and endpoint of bilingual development are multiple, usually not dichotomous, and vary over time. In today’s world of patchwork families, for example, it is easily conceivable for a child to be exposed to two languages from birth for a limited period and then to live in a monolingual (two- or one-parent) household which later on becomes bilingual again (not necessarily in the same language combination as before). Increased employment mobility may also lead to varying combinations of home and school/work languages over the lifetime. The same issues exist for the learning context, which is not necessarily constant. Many bilinguals who begin the learning process in an instructed setting will, at some point, also benefit from naturalistic acquisition, for example in a Study Abroad experience. Similarly there are many migrants who largely acquire the L2 from the input they receive in their everyday environment, but who decide (or are obliged by governmental rules) to also attend some language classes.

Such ‘hybrid’ cases not only beg the question of how they should be labeled, but also whether they can and should be included in experimental research investigating the impact of bilingualism. Should they be treated as ‘anomalies’ which will only

obscure the picture, or be considered important to the understanding of the overall process? For studies opting for the former it will quickly become evident that a purist approach to participant selection will mean that only a very small subpopulation of learners remain available for study. In turn this raises questions on whether the results gained from such a limited subset can be generalized to the wider reality where factors are more varied and may also interact with each other in unpredictable ways.

Another issue which has proven a stumbling block within research into bilingualism is that limiting a study to a single factor, a limited aspect of language, and/or one particular elicitation technique, can be misleading. Various factors may impact differently on various types of knowledge. Age at onset, for example, may have a stronger impact on pronunciation skills than on lexical acquisition, while context of learning may manifest itself more strongly on one type of task than on another. This means that studies adopting a narrow research design may only arrive at a very partial picture.

1.2 The Monolingual Baseline

A further important question is what constitutes an appropriate native baseline. In most studies, this consists of (predominantly) monolingual natives. Let us set aside, for the moment, the question of whether bilinguals and monolinguals are underlyingly, 'fundamentally', or qualitatively different from each other with respect to issues such as grammatical representations, lexicon and phonology. While these issues are highly controversial, there is no doubt that all bilinguals face a challenge from which all monolinguals are spared: In learning, using, processing and maintaining language, bilinguals have to keep separate two (or more) knowledge systems which share essentially the same function and structure but which are arbitrarily different in their inventories. All languages have phonology, morphology, syntax, a lexicon and pragmatics, but they differ with respect to the phonemes, morphemes, word order rules, lexical items, etc. which they use. It is therefore virtually inevitable that bilinguals will, on occasion, produce mixed variants or code-switch, in particular when other concurrent cognitive demands are high and/or attention to the message is low due to fatigue or distraction. They will also almost invariably be slower in accessing information, simply because their volume of knowledge is larger. While such phenomena are absent in the speech of monolinguals, their impact on the behavior or processing of bilinguals is not indicative of a 'deviation' in the knowledge underpinning language production and understanding. This presents a significant challenge to investigations of bilingual acquisition and its (potential) limitations, since it is hard to separate out interference between the two linguistic systems at the level of production or processing from non-nativeness at the level of representation. Any linguistic task is more complex and challenging for bilingual speakers than for monolinguals, since they not only have to solve the task in one language but suppress performing it in the other (Hopp

and Schmid 2013; Schmid 2014). Nevertheless, the baseline for investigations of ultimate L2 attainment and its potential limitations are almost invariably (predominantly) monolingual speakers.

We would like to challenge the assumption that monolinguals are the appropriate point of reference. In our view, in order to establish whether there are indeed linguistic phenomena that persistently elude full L2 acquisition, it is necessary to determine that the problem is not merely a consequence of language interference. To do so, a different population needs to be invoked: Speakers who have learned a first language as their only native language from birth, but who experience comparable competition from a different language in the same way the late learners do. In other words, we argue for the inclusion of a reference population of speakers who have experienced the linguistic development known as language dominance reversal, or *first language attrition*.

L1 attrition, as used here, refers to a gradual decline in proficiency in a native language among bilingual migrants. Attrition is the other side of bilingual development; as a speaker uses his or her second language frequently and becomes proficient in it, some aspects of the first language can become subject to L2 influence or deteriorate. The same factors that play a role in L2 acquisition appear to mediate L1 attrition, for example exposure and use (e.g. Hulsen 2000; Schmid 2007; Schmid and Dusseldorp 2010), attitude and motivation (Ben-Rafael and Schmid 2007; Schmid 2002) or aptitude (Bylund 2008; Bylund and Ramírez-Galan 2014). For this reason, comparisons of L2 learners and L1 attriters can potentially shed light on central questions about the nature of bilingual knowledge and processing (Schmid 2009).

The choice of a baseline of attrited native speakers, however, does pose a number of practical problems, since these speakers invariably reside in a different linguistic environment than the (immersed) L2 learners and/or monolingual speakers with whom they are compared. This means that data will need to be collected at a variety of sites and across different countries, which may lead to problems of the compatibility of both hardware and software in data acquisition, the comparability of the background of the speakers and other factors important for a clean comparison. These problems are addressed in the subsequent chapters.

A preliminary, important consideration, however, concerns the choice of linguistic variable or variables for investigation.

1.3 Choosing Crosslinguistic Variables for Testing

The controversy around the mechanisms underpinning first versus second language acquisition usually hinges on particular linguistic features. It appears to be uncontroversial that second-language learners can become native-like with respect to some aspects of language. Semantics is often cited as a case in point. For structural or grammatical features of the language, there is also evidence that advanced late L2 learners can develop native-like processing at least for some

features, in particular the ability to detect errors that would be ungrammatical in the participants' L1 as well as in the L2 such as phrase-structure rules (e.g. Rossi et al. 2006), violations of word order (Bowden et al. 2013) or morphology (e.g. Rossi et al. 2006). For other grammatical phenomena, however, questions of ultimate attainment are more difficult to resolve. Opinions are divided on the extent to which it is possible for L2 learners to acquire linguistic features involved with some movement operations (e.g. *wh*-movement, see Belikova and White 2009 vs. Hawkins and Hattori 2006 or scrambling, see Hopp 2006) or inflectional features such as case, tense and gender marking (see Hopp 2010 for an overview), in particular if they are not represented in the learners' L1. Grammatical gender is another feature that has been widely and controversially discussed in this context. Given the controversies on representational or processing differences between L1 and L2, gender is a fruitful area for bilingualism research, since native-like knowledge and processing of this feature necessitates coordination of information from different linguistic levels, gender being encoded as part of the lemmatic information in the mental lexicon on the one hand and triggering concord across the items in the phrase or sentence on the other.

1.3.1 Grammatical Gender as a Challenge to Bilingual Development

Many languages assign nouns to different categories, triggering agreement on other elements within or outside the NP (determiners, adjectives, quantifiers, verbs, anaphoric pronouns, etc.); this phenomenon is commonly referred to as 'gender concord'.[1] While in some languages noun classification is to some degree transparent, based on semantic or phonological criteria (e.g. Spanish, see Franceschina 2005; Foucart 2008), in others, among them most Germanic languages, it appears arbitrary and largely unpredictable. The acquisition of Germanic gender systems appears relatively unproblematic for children (gender concord has been observed to be in place around age 3;3–3;6, see Bewer 2004; Mills 1986; Müller 1994 for German and Gillis and De Houwer 1998 for Dutch; although occasional variability may be witnessed in Dutch until around age 6 according to Blom et al. 2008). However, it is notoriously and persistently difficult for L2 learners (e.g. Fries 2001; Rogers 1987; Sabourin 2003; Sabourin and Stowe 2008) but there is to date little evidence of grammatical gender systems being adversely affected by language attrition in late bilinguals: Scherag et al. (2004) find that long-term migrants perform in a perfectly native-like fashion on an experiment using a gender-priming paradigm. Gender marking appears to be remarkably stable in free production as well (e.g. Schmid 2002, 2014).

[1]The present discussion is confined to linguistic items which refer to inanimate objects which do not have a natural gender.

There is furthermore considerable converging evidence that gender concord facilitates lexical access in L1 processing. Priming experiments consistently show a gain in response times (RTs) if the target item is preceded by a congruently marked article or adjective (Bates et al. 1996; Grosjean et al. 1994), and picture naming is facilitated when a picture of an object is presented together with a (semantically unrelated) distracter word which has the same gender as the object to be named (Schiller and Caramazza 2003; Schriefers 1993). These findings are taken to indicate that syntactic features such as gender are activated in bare lemma selection, and that gender concord reduces the 'pool' of possible candidates for lexical access in native speakers.

To what degree L2 speakers can exploit gender marking in this way is controversial. A replication of the experiment conducted by Grosjean et al. (1994) suggested that early L2 learners whose L1 does not instantiate gender have facilitation and inhibition effects that are similar to those of L1 speakers, while late L2 learners do not have the same sensitivity to congruent and incongruent gender marking (Guillelmon and Grosjean 2001). L2 speakers also appear to rely on phonological clues when asked to name the gender of bare nouns presented to them, while for native speakers phonological gender-transparency does not influence RTs (Bordag et al. 2006; Hohlfeld 2006). These findings suggest that late L2 learners may indeed "fundamentally differ from monolinguals in important respects" and that "learners' lexical representations of gender are often weak, unstable or even incorrect" (Lemhöfer et al. 2008: 327).

Where native perception of gender concord is concerned, there is solid evidence for strong and reliable responses to concord violation in the EEG signal from a range of languages including Dutch (Hagoort and Brown 2000; Sabourin and Stowe 2008), German (Davidson and Indefrey 2009; Gunter et al. 2000), French (Foucart and Frenck-Mestre 2012), Hebrew (Deutsch and Bentin 2001), Spanish (Barber and Carreiras 2005) and Italian (Molinaro et al. 2008). Among L2 learners, on the other hand, the available evidence is again less than clear, with some studies finding responses that are comparable to those of natives (e.g. Alemán Bañón et al. 2014) while others find differences (Foucart and Frenck-Mestre 2012; Sabourin and Stowe 2008, for more detail see Chap. 6).

1.3.2 Grammatical Gender in German and Dutch

Examining gender also allows us to investigate a factor which has not received much attention in the literature, namely how and why the gender systems of some languages may be more or less difficult to master in L2 acquisition. While the impact of similarities between L1 and L2 in this respect have frequently been studied, little attention has been paid to inherent properties of the L2 system as such. For example, the fact that so many studies of the L2 acquisition of gender focus on Spanish, a language in which the gender of 90 % of all items is predictable based on the noun coda (Franceschina 2005) may have led to an overestimation of

Table 1.1 The inflectional paradigm of German NPs (*der große Berg* (masc.) 'the big mountain', *die große Strasse* (fem.) 'the big street', *das große Tal* (neut.), 'the big valley')

Case	Masculine		Feminine		Neuter	
	Definite	Indefinite	Definite	Indefinite	Definite	Indefinite
N	**der** groß**e** Berg	**ein** groß**er** Berg	**die** groß**e** Strasse	**eine** groß**e** Strasse	**das** groß**e** Tal	**ein** groß**es** Tal
G	**des** groß**en** Berg**es**	**eines** groß**en** Berg**es**	**der** groß**en** Strasse	**einer** groß**en** Strasse	**des** groß**en** Tal**es**	**eines** groß**en** Tal**es**
D	**dem** groß**en** Berg	**einem** groß**en** Berg	**der** groß**en** Strasse	**einer** groß**en** Strasse	**dem** groß**en** Tal	**einem** groß**en** Tal
A	**den** groß**en** Berg	**einen** groß**en** Berg	**die** groß**e** Strasse	**eine** groß**e** Strasse	**das** groß**e** Tal	**ein** groß**es** Tal

the ability of L2 learners to implement gender as a grammatical feature. Similarly, questions of how many genders there are in a language, how varied and salient the inflectional paradigm is (consider French where a large proportion of gender agreement is phonologically obscured and only visible in the written language), and so forth, have rarely been considered. In this respect, German and Dutch (two closely related languages) present an interesting contrast. While gender is an inherent property of all German and Dutch nouns, it is not directly expressed on the nouns and only manifests itself by agreement on other items within the noun phrase: In this sense both languages are non-transparent. German furthermore has three grammatical genders (masculine, feminine and neuter) while Dutch has only two (common and neuter). Common gender in Dutch is a conflation of the historical categories of masculine and feminine and comprises ca. 80 % of all nouns, making it a clear choice for a default. The German inflectional paradigm (see Table 1.1) is also much richer than the Dutch one (see Table 1.2). German marks gender, definiteness, and case by inflection on determiner and adjective, while Dutch has only two forms of the definite article (*de* and *het* in the singular, *de* in the plural), one form of the indefinite article (*een*), and two forms of inflection on the adjective (Ø in indefinite singular neuter NPs, *-e* elsewhere). Neither Dutch nor German mark gender morphologically in the plural.

The two systems thus pose differential challenges to L2 learners: On the one hand, German gender is characterized by a large number of syncretisms and ambiguous morphological forms (for example, masculine and neuter dative forms are identical), which may lead to confusion and, if used incorrectly, to ambiguous or unintended meanings. On the other hand, these inflectional markers are more

Table 1.2 The inflectional paradigm of Dutch NPs (*de grote berg* (common) 'the big mountain', *het grote dal* (neut.), 'the big valley')

Common		Neuter	
Definite	Indefinite	Definite	Indefinite
de grote berg	**een** grote berg	**het** grote dal	**een** groot**Ø** dal

salient than concord in Dutch, since in Dutch the adjectival inflection is almost always the same form regardless of the gender of the noun, leaving the singular determiner as the sole cue to gender. Comparing the L2 acquisition and L1 attrition of gender in these two languages thus presents the potential to assess the impact of a relatively minimal and straightforward agreement contrast which does little to contribute to reference resolution (in Dutch) versus a more elaborate system of agreement markers which can be assumed to be more noticeable due to its interaction with other grammatical features. This makes this particular linguistic feature an excellent target variable for investigations of bilingual development, contrasting its acquisition by L2 speakers from a range of linguistic backgrounds which do or do not mark gender morphologically and its maintenance among L1 speakers in an attrition setting with processing among monolingual native speakers.

1.4 Methodological Challenges

The various considerations sketched above, and others that relate to them (e.g. factors such as the amount of input an individual learner receives, the typological distance of the languages involved, levels of learner aptitude, attitude and motivation), present a formidable methodological challenge to the field of bilingualism research. In order to eventually arrive at a more comprehensive account of a phenomenon which is shared by substantially more than half of the world's population and is increasingly identified as both a resource and a problem for modern societies, the present text outlines a number of desiderata for a large scale study that deals with some of these difficulties.

First and foremost, we argue that investigations of bilingualism should always address more than one aspect of a speaker's linguistic repertoire. These aspects include (but are not limited to):

- formal knowledge which is open to introspection as assessed, for example, by tasks which involve grammaticality or acceptability judgments
- implicit knowledge of the same structures, for example responses evident in online grammatical processing which takes place before explicit knowledge is accessed
- language use under naturalistic conditions, i.e., under constraints of real-time interaction and of cognitive limitations due to other concurrent processing demands normally present in comprehension and production.

Second, the investigation should comprise different populations of bilingual speakers. It should not only compare L2 learners with L1 monolinguals and attriters, but should also include speakers whose alternative language does or does not encode the feature under investigation, in order to allow conclusions with respect to the role of crosslinguistic facilitation and interference.

Since no single study can possibly either address or control for all of the internal and external factors surrounding the acquisition and attrition process, nor cover all

of these elicitation methods and different types of bilingual populations, converging evidence from different studies is necessary in order to eventually gain insight into what impact these factors have. This includes studies investigating bilinguals with a range of ages of onset and proficiency levels, coming from different learning contexts and language backgrounds. For the converging evidence to be maximally informative, a carefully planned set of studies is best geared to produce data with maximal relevance to the overall goal, including similar tasks across target languages and identical tasks across bilingual groups. One of the issues to be dealt with in such a converging approach is the fact that it inevitably implies collecting data at many different sites which may not be directly comparable due to different equipment, design and setup.

1.5 A Sample Study

The present text provides a description of how to set up and carry out a study which involves collecting data at different sites, with different methodologies, and testing populations varying along multiple dimensions. As an example, it documents the considerations and measures adopted in a large-scale investigation of more than 300 monolinguals and bilinguals of varying ages of acquisition (AoA), language backgrounds and combinations of the two. Two closely related target languages, Dutch and German, were investigated, not only as the L2 of learners from different native language backgrounds, but also as the L1 of long-term migrants who use another language predominantly in their daily lives and have done so for many years (L1 attriters). We recruited populations of L2 learners with native languages that were typologically distant from Dutch and German. Approximately half of these were native speakers of languages that also encode gender (Polish and Russian), while the other half came from a linguistic background that lacked this feature entirely (comprising languages such as Turkish, Farsi and Chinese). It would have been ideal for the comparison to recruit attriting populations of Dutch and German in the opposite language settings (i.e., in Poland, Russia, Turkey etc.). This proved to be impossible for practical reasons due to the limited emigration of monolingual German and Dutch speakers to these countries, and we therefore opted to study the L1 of these populations in English-speaking settings in North America. A (predominantly if not perfectly) monolingual baseline was also included.

The linguistic feature we were predominantly interested in was gender concord, as this grammatical marker presents a particular challenge for bilingual language acquisition and maintenance and differs in theoretically interesting ways between the two languages under investigation (see Sect. 1.3.2). In order to gain comprehensive insight into how gender concord is accessed and used in these populations, a variety of elicitation techniques were used. Offline grammaticality and gender assignment tasks (see online supplementary material) provided information on the participants' metalinguistic knowledge of gender concord rules. A free speech task was used to tap into the ability to apply these rules under the cognitive demands of real-time

language production. Additionally this task allowed an investigation of the wider skills involved in naturalistic language production in these populations. Two neurocognitive methods, EEG and eyetracking, allowed us to gain insight respectively into online sensitivity to grammatical violation and the use of gender information to facilitate language comprehension through creation of expectations. In this way we hoped to obtain a comprehensive picture of which aspects of this grammatical feature were more or less challenging for the different populations. In particular the latter two types of experiments, tapping into neurocognitive processes through EEG and eyetracking, presented formidable challenges for the acquisition of data sets at different sites that would be comparable with each other (see Chaps. 3, 5 and 6 for more detail). Since these kinds of experimental approaches are much more problematic for multi-site studies than, for instance, studies measuring reaction times we present the outcome of our considerations here in some detail.

A substantial amount of background information collected from our participants allowed us to gain more insight into factors facilitating or impeding successful acquisition and maintenance. The results from the various experimental approaches are reported elsewhere; in this text, we focus on methodology and background and their relationship to designing a large scale study of this kind. The aim of the present text is to discuss the challenges as we perceived them and to make information available to the wider linguistic community about how we dealt with them, in the hope that our experience may be helpful to others and eventually contribute to general research standards which will facilitate cross-study comparisons.

1.6 Overview of the Book

The present volume is structured into the following sections: Chap. 2 presents an overview of background variables and predictor factors that are important for investigations of bilingual development, and offers some suggestions as to how to measure and assess these factors. Chapter 3 addresses in more detail the practical challenges involved with testing at different sites. Chapters 4–6 each address a particular method for linguistic research: Chap. 4 concerns itself with the elicitation, collection and analysis of naturalistic speech; Chap. 5 is dedicated to the use of eye-tracking, in particular within the paradigm known as Visual World, while Chap. 6 treats the use of EEG for research on grammatical violations.

References

Alemán Bañón, J.A., R. Fiorentino, and A. Gabriele. 2014. Morphosyntactic processing in advanced second language (L2) learners: An event-related potential investigation of the effects of L1–L2 similarity and structural distance. *Second Language Research* 30(3): 275–306.

Barber, H., and M. Carreiras. 2005. Grammatical gender and number agreement in Spanish: An ERP comparison. *Journal of Cognitive Neuroscience* 17(1): 137–153.

Bates, E., A. Devescovi, A. Hernandez, and L. Pizzamiglio. 1996. Gender priming in Italian. *Perception and Psychophysics* 58(7): 992–1004.

Belikova, A., and L. White. 2009. Evidence for the fundamental difference hypothesis or not? *Studies in Second Language Acquisition* 31(02): 199–223.

Ben-Rafael, M., and M.S. Schmid. 2007. Language attrition. In *Theoretical perspectives*, ed. B. Köpke, M.S. Schmid, M. Keijzer, and S. Dostert, 205–226. Amsterdam/Philadelphia: John Benjamins.

Bewer, F. 2004. Der Erwerb des Artikels als Genus-Anzeiger im deutschen Erstspracherwerb. *ZAS Papers in Linguistics* 33: 87–140.

Bley-Vroman, R. 2009. The evolving context of the fundamental difference hypothesis. *Studies in Second Language Acquisition* 31(2): 175–198.

Blom, E., D. Polišenská, and F. Weerman. 2008. Articles, adjectives and age of onset: the acquisition of Dutch grammatical gender. *Second Language Research* 24(3): 297–331.

Bordag, D., A. Opitz, and T. Pechmann. 2006. Gender processing in first and second languages: The role of noun termination. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 32(5): 1090.

Bowden, H.W., K. Steinhauer, C. Sanz, and M.T. Ullman. 2013. Native-like brain processing of syntax can be attained by university foreign language learners. *Neuropsychologia* 51(13): 2492–2511.

Bylund Spångberg, Emanuel. 2008. *Age differences in first language*. PhD thesis, Stockholm University.

Bylund, E., and P. Ramírez-Galan. 2014. Language aptitude in first language attrition: A study on late Spanish Swedish bilinguals. Under review (Applied Linguistics).

Davidson, D.J., and P. Indefrey. 2009. An event-related potential study on changes of violation and error responses during morphosyntactic learning. *Journal of Cognitive Neuroscience* 21(3): 433–446.

Deutsch, A., and S. Bentin. 2001. Syntactic and semantic factors in processing gender agreement in Hebrew: Evidence from ERPs and eye movements. *Journal of Memory and Language* 45(2): 200–224.

Foucart, A. 2008. *Grammatical gender processing in French as a first and a second language*. PhD dissertation Université Aix-Marseille I/University of Edinburgh.

Foucart, A., and C. Frenck-Mestre. 2012. Can late L2 learners acquire new grammatical features? Evidence from ERPs and eye-tracking. *Journal of Memory and Language* 66(1): 226–248.

Franceschina, F. 2005. *Fossilized second language grammars: The acquisition of grammatical gender*, Vol 38. Amsterdam: John Benjamins Publishing.

Fries, N. 2001. Ist Deutsch eine schwere Sprache? Am Beispiel des Genus. In *Die deutsche Sprache in der Gegenwart*, ed. S. Schierholz, 131–146. Berlin: Peter Lang.

Gillis, S., and A. De Houwer. 1998. *The acquisition of Dutch*. Amsterdam/Philadelphia: John Benjamins.

Grosjean, F., J.Y. Dommergues, E. Cornu, D. Guillelmon, and C. Besson. 1994. The gender-marking effect in spoken word recognition. *Perception and Psychophysics* 56(5): 590–598.

Guillelmon, D., and F. Grosjean. 2001. The gender marking effect in spoken word recognition: The case of bilinguals. *Memory & Cognition* 29(3): 503–511.

Gunter, T., A. Friederici, and H. Schriefers. 2000. Syntactic gender and semantic expectancy: ERPs reveal early autonomy and late interaction. *Cognitive Neuroscience, Journal of* 12(4): 556–568.

Hagoort, P., and C.M. Brown. 2000. ERP effects of listening to speech compared to reading: the P600/SPS to syntactic violations in spoken sentences and rapid serial visual presentation. *Neuropsychologia* 38(11): 1531–1549.

Hawkins, R., and H. Hattori. 2006. Interpretation of English multiple wh-questions by Japanese speakers: A missing uninterpretable feature account. *Second Language Research* 22(3): 269–301.

Hohlfeld, A. 2006. Accessing grammatical gender in German: The impact of gender-marking regularities. *Applied Psycholinguistics* 27:127–142.

Hopp, H. 2006. Syntactic features and reanalysis in near-native processing. *Second Language Research* 22(3): 369–397.

Hopp, H. 2010. Ultimate attainment in L2 inflection: Performance similarities between non-native and native speakers. *Lingua* 120(4): 901–931.

Hulsen, M. 2000. Language Loss and Language Processing. In *Three Generations of Dutch Migrants in New Zealand*. PhD dissertation, Nijmegen: Katholike Universiteit Nijmegen.

Lemhöfer, K., T. Dijkstra, H. Schriefers, R.H. Baayen, J. Grainger, and P. Zwitserlood. 2008. Native language influences on word recognition in a second language: A megastudy. *Journal of Experimental Psychology. Learning, Memory, and Cognition* 34(1): 12.

MacWhinney, B. 2012. The logic of the unified model. In *The Routledge handbook of second language acquisition*, ed. S. Gass, and A. Mackey, 211–227. Abbingdon: Routledge.

Mills, A.E. 1986. *The Acquisition of Gender. A study of English and German*. Berlin: Springer.

Molinaro, N., F. Vespignani, and R. Job. 2008. A deeper reanalysis of a superficial feature: An ERP study on agreement violations. *Brain Research* 1228: 161–176.

Müller, N. 1994. Gender and number agreement within DP. In *Bilingual First Language Acquisition: French and German grammatical development*, ed. J.M. Meisel, 53–88. Amsterdam: John Benjamins.

Paradis, M. 2009. *Declarative and procedural determinants of second languages*. Amsterdam: John Benjamins Publishing.

Rogers, M. 1987. Learners' difficulties with grammatical gender in German as a foreign language. *Applied Linguistics* 8: 48–74.

Rossi, S., M. Gugler, A. Friederici, and A. Hahne. 2006. The impact of proficiency on syntactic second-language processing of German and Italian: Evidence from event-related potentials. *Cognitive Neuroscience* 18(12): 2030–2048.

Sabourin, L. 2003. *Grammatical Gender and Second Language Processing*. PhD dissertation, University of Groningen.

Sabourin, L., and L.A. Stowe. 2008. Second language processing: when are first and second languages processed similarly? *Second Language Research* 24(3): 397–430.

Scherag, A., L. Demuth, F. Rösler, H.J. Neville, and B. Röder. 2004. The effects of late acquisition of L2 and the consequences of immigration on L1 for semantic and morpho-syntactic language aspects. *Cognition* 93(3): B97–B108.

Schiller, N.O., and A. Caramazza. 2003. Grammatical feature selection in noun phrase production: Evidence from German and Dutch. *Journal of Memory and Language* 48(1): 169–194.

Schmid, M.S. 2002. *First language attrition, use and maintenance: The case of German Jews in Anglophone countries*. Amsterdam: John Benjamins Publishing.

Schmid, M.S. 2007. The role of L1 use for L1 attrition. In *Language attrition: Theoretical perspectives*, eds. B. Köpke, M.S. Schmid, M. Keijzer and S. Dostert, 135–153. Amsterdam: John Benjamins.

Schmid, M.S. 2009. On L1 attrition and the linguistic system. *Eurosla Yearbook* 9(1): 212–244.

Schmid, M.S., and E. Dusseldorp. 2010. Quantitative analyses in a multivariate study of language attrition: the impact of extralinguistic factors. *Second Language Research* 26(1): 125–160.

Schmid, Monika S. 2013. First language attrition: state of the discipline and future directions. *Linguistic Approaches to Bilingualism* 3(1): 97–116.

Schmid, Monika S. 2014. The debate on maturational constraints in bilingual development: A perspective from first language attrition. *Language acquisition* 21(4): 386–410.

Schriefers, H. 1993. Syntactic processes in the production of noun phrases. *Journal of Experimental Psychology. Learning, Memory, and Cognition* 19(4): 841–850.

Chapter 2
Multi-factorial Studies: Populations and Linguistic Features

Monika S. Schmid

Abstract It was suggested in the introduction to this volume that cumulative evidence from studies investigating populations of learners from varying linguistic backgrounds, under different learning contexts, and with a range of experimental designs is necessary in order to gain further insight into fundamental questions of bilingual development—such as whether first and second language learning are qualitatively different from each other, or whether cognitive maturation specifically and independently affects language learning. Such a meta-approach, however, is easily compromised, as what may seem to be relatively minor differences and adjustments in participant selection, data acquisition, coding and analysis may eventually make it impossible to compare the findings from one study to that of another, or lead to conflicting findings. This chapter discusses the issue of what factors characterizing the populations being tested should and can be documented. We start with a discussion of how the lack of such documentation and differences in analysis have muddied the waters in the past.

Keywords Sociolinguistic and personal background factors · Second language acquisition · First language attrition

2.1 Cross-Study Variability and the Importance of Participant Documentation

"Even a supposedly clean study is prone to alternative interpretations of data" (Bialystok and Hakuta 1994, p. 69). This statement refers to the seminal study by Johnson and Newport (1989, henceforth J&N) which was long accepted as providing conclusive evidence for the existence of a maturational constraint on

Electronic supplementary material The online version of this article (doi:10.1007/978-3-319-11529-0_2) contains supplementary material, which is available to authorized users.

M.S. Schmid et al., *Designing Research on Bilingual Development*,
SpringerBriefs in Linguistics, DOI 10.1007/978-3-319-11529-0_2

second-language learning (the so-called ‘Critical Period’). J&N investigated a population of 46 Chinese and Korean learners of English with varying ages of acquisition (AoAs) by means of an extensive offline auditory grammaticality judgment task comprising 276 items. They found that the scores achieved on this task correlated strongly with AoA for the segment of their population who had acquired English before age 15 but not for those who had learned it after that age. This discontinuity was interpreted as evidence for a qualitative change of the learning mechanisms linked to biological development: “[L]anguage learning ability slowly declines as the human matures and plateaus at a low level after puberty” (p. 90).

Later investigations, however, are divided on whether or not to accept J&N’s conclusions. On the one hand, a re-analysis of the original data by Bialystok and Hakuta (1994) found that when the cutoff point between the younger and the older population was moved from 15 to 20, age of learning remained a significant predictor in both groups and the apparent discontinuity found in the original study disappeared. Bialystok and Hakuta further point out that the age-related decline observed across the sample is not necessarily the result of the deterioration of a *language-specific* acquisition ability, but may be due to other factors such as age at testing, different language learning circumstances in the different AoA segments and so on (p. 70f.). Similarly, Elman et al. (1996) propose different statistical ways of analyzing J&N’s data and suggest that ‘a simple linear dynamical alternative is available to account for the same nonlinear outcomes’ (p. 188).

An additional issue is that the population involved is inadequately described. It seems clear that other factors than AoA must be involved given the different conclusions that replications of the study have found: Birdsong and Molis (2001) find the exact opposite pattern of AoA effects among 61 native Spanish learners of English (no correlation between age of onset and score on the GJT for learners who were below 15 when they began the acquisition process but a strong correlation for learners beyond that age). Conversely, DeKeyser (2000), investigating 57 Hungarian learners of English, and DeKeyser et al. (2010), investigating 76 Russian learners of English and 64 Russian learners of Hebrew, found discontinuities in all populations which were similar to those discovered by J&N. Without detailed characterizations of the populations it is difficult to reconcile these results.

Given the large range of factors that impact on L2 development, it is not surprising that findings from different studies are often inconsistent or controversial. Inconsistencies such as the ones sketched above, however, severely limit our ability to come to hard conclusions about the language learning process and what constrains it. Lamentably, there is no overall consensus on how populations for linguistic investigations should be selected, and what factors it is vital to document for comparison purposes. This chapter will make an attempt to identify a number of these factors and suggest ways in which they may be assessed in order to ensure that populations and, eventually, studies, can be characterized more comprehensively. We hope that this may allow explanations of contradictory findings in terms of relevant group differences on the basis of cross-study comparisons.

The list of factors necessarily comprises biodata such as age, AoA, gender, education, learning context, SES and so forth, which are comparatively easy to elicit and record. Second, there are variable individual traits, potentials and limitations which, while not in themselves language-specific, may affect or impair performance on learning and on particular linguistic tasks and thus serve to distort the picture. This sort of variable comprises matters such as handedness; any potential impairments to sensory modalities: vision (including color blindness) or hearing; and cognitive factors like Working Memory (WM) and/or intelligence. In this context, we need to acknowledge the difficulties posed in measuring and comparing some of these factors across populations with different first languages. Lastly, the mere fact of knowing two languages will have an effect on knowledge and use of both languages, producing differences between monolinguals and bilinguals.

In some cases, information about the speakers should be collected beforehand in order to determine whether there are reasons to exclude them from participating in the study. For example, many executive control tasks such as the Stroop task (Stroop 1935), Go-No-Go tasks or language switching tasks use different colors as cues to which response should be given. Participants who have any kind of impairment to color perception will almost certainly provide different responses than participants with unimpaired perception. For tasks which rely on fast response times, participants above a certain age may be too slow, adding undue noise to the data being analyzed, and the researcher may therefore decide to establish an upper age limit for the population or exclude on an individual basis determined by the population distribution. How the criteria discussed here will be applied in any individual investigation will depend on the research design and research questions, so the procedures we outline here serve merely as suggestions.

Most of the factors discussed below, however, do not determine whether participants are included, but must be documented in order to be used in characterizing the populations. This relates to factors which may facilitate or constrain performance on certain tasks, such as level of education or language proficiency. For many of the more traditional statistical approaches, it is desirable that the different populations are as similar as possible with respect to these factors. Some more novel approaches, such as mixed effects and multiple regression models are better equipped for dealing with variability in the range of predictors and additionally do not force the use of artificial groupings within linear factors like AoA, but similar age ranges remain important for group comparisons.

In order to obtain the desired range of participants, balance criteria such as education level and gender, and apply exclusion criteria, it is a good idea to elicit or at least estimate some of these factors from potential participants before they are actually recruited for the study. This is comparatively straightforward for biographical criteria, such as age of learning, age at testing, length of exposure to the target language, and so on, which can relatively easily be collected by an online participant questionnaire. For other factors, such as handedness, the assessment is somewhat more complex, whereas measuring predictors such as language proficiency is invariably rather time-consuming. It is important that the pre-assessment

should strike the right balance here: If it becomes too lengthy, participants will be put off before they ever attend a testing session; if, on the other hand, it is too short and superficial, participants who do not fit the testing criteria (for example because their proficiency level is too low) will be included but their data may not be usable, and this may be expensive in terms of both laboratory costs and working time—not to mention unfair towards the participants.

2.2 Exclusion Criteria

Since we are usually keen to have populations as large as possible, excluding participants who have volunteered to take part in the study may not seem desirable. However, a number of participation criteria need to be established since otherwise results may be skewed, flawed or obscured. These exclusion criteria fall broadly into two categories: Factors that are connected with experience with additional languages not being tested, and physical and health factors that may affect the performance of the participant.

First, it is important to establish each individual's **language learning history**. For investigations that wish to study a particular process of bilingual acquisition it is often important that none of the participants has had substantial experience with languages other than those under investigation, in particular 1) outside the range of age of acquisition for which they are being tested and/or 2) containing a linguistic characteristic similar to the one under investigation (e.g. tone or gender), as this may facilitate its acquisition in the third language. This normally excludes speakers who grew up in bilingual families, have spent extended periods of time in third countries (these may include countries or regions where different varieties of the target language are spoken, for example Austria or Switzerland vs. Germany or Francophone regions in Canada or Africa vs. France) or otherwise achieved advanced proficiency in another language. We therefore suggest that in the course of the pre-screening, a short language learning questionnaire should be administered (see online supplementary material), and that it should be carefully considered to what extent any prior linguistic experiences may obscure or influence performance. Note that similar criteria should be applied to the control group, as well. One frequently finds investigations of, for example, native speakers of American English acquiring Spanish, where the reference population is comprised of speakers born in Latin American countries and residing in the US. While this is certainly a convenient and inexpensive way of recruiting a control group, such speakers may have experienced language dominance reversal and/or L1 attrition, particularly if they emigrated to the US early and were educated there, which may well distort the findings (Dussias 2004).

Second, it is advisable to establish whether participants are affected by a **medical condition** or **disorder** which might impact their performance. These include conditions that will impair the perception of or response to the stimuli, such as visual or hearing impairments (unless corrected), dyslexia, speech disorders (such as

stuttering) and color blindness. A number of neurological conditions (such as epilepsy, strokes or brain lesions) or neurodegenerative disorders may also result in linguistic difficulties, and the use of various medications, drugs or alcohol may interfere with task performance. It is wise not to include participants who are affected by any of these.

Lastly, **handedness** can be an important factor. For studies which measure response time by means of pressing a particular key on a computer keyboard or on a Serial Response Box, left-handed participants should be tested by means of a setup that is the reverse of that of right-handed ones. Although handedness does not strongly correlate with language lateralization, it does provide an indication and can muddy the waters when brain measures are used, since in left-handers there is a 20 % chance that the functional organization of the hemisphere is inverted (McKeever et al. 1995). For this reason, it is recommended that the population should comprise only right-handed individuals in studies employing neurological and neurolinguistic measures, such as EEG or MRI readings. Handedness, however, is not a dichotomous concept which neatly divides the population into left- and right-handed individuals, but rather a gradient phenomenon with more or less strongly pronounced levels. It has been established that the chance of right hemisphere dominance increases as the degree of left handedness increases (Knecht et al. 2000). For the purpose of our investigation we relied on an abbreviated version of the Edinburgh Handedness Inventory (Oldfield 1971; the version used in our study was part of the prescreening questionnaire which can be accessed in the online supplementary material, and excluded any participants who did not clearly emerge as right-handed.

2.3 Personal Variables

Many individual factors may have an impact on proficiency in both L1 and L2. Among these are biographical factors which are the outcome of lifetime events and impact indirectly on language development and processing. Others concern the cognitive skills that influence an individual's ability to learn and process language. Factors which depend on personal habits and/or attitudes are often directly linked to language proficiency. Not all studies will need to include all of these factors to the same level of detail, but in designing the study it is important to consider which of them should be assessed so that their impact on various populations can be examined. Even when a study is not directly concerned with some of these factors, it is recommended to include them to facilitate cross-study comparisons in the future. Furthermore, some factors may be used to establish the exclusion criteria discussed above.

2.3.1 Biographical Data

The factors treated in this section are characteristics of all participants, whether mono- or bilingual, which may have an impact on their performance on various tasks and should therefore be sampled similarly in the different segments of the population. First, and probably most straightforward, the **sex** of participants should be matched across populations. The debate on the impact of sex for language acquisition is too wide-ranging to be treated in detail here (as is the relatively recent debate on whether biological sex is a dichotomous factor; see for example Hall 2008). However, there is ample evidence to suggest that, both in instructed and in immersed language learning, females may have a degree of advantage over males (e.g. Pavlenko et al. 2001). This debate goes back to the age-old nature versus nurture controversy which asks to what extent such differences are due merely to socially constructed realities and expectations, or whether there is some kind of impact of differential (probably hormonally modulated) use of neurocognitive resources by the two genders that may make women more effective language learners (see the discussion by Ullman 2004). Given, however, that many studies of (first- and second-) language processing and production have found differences between male and female populations, it is wise to aim for equal representation of both genders. Should this not be possible, proportions of women and men should at least be approximately the same in all populations.

Second, the **age** of participants[1] at the time of testing is of importance. Performance on a variety of tasks fluctuates and changes throughout the lifespan, so that care should be taken to ensure that all populations are approximately age-matched. Furthermore, many recent studies suggest that age-related declines in cognitive skill may to some extent be attenuated in bilingual populations (e.g. Abutalebi et al. 2014; Bialystok et al. 2008; Gold et al. 2013). This means that if the populations comprise large numbers of individuals at higher age ranges, differences that are due to bilingual proficiency and the end state of long-term adult L2 learning (frequently referred to as ultimate attainment) may be confounded with differences contingent on the protective influence of long-term bilingualism on cognitive function. It is therefore strongly advised that a) not only the mean age but also the age range across populations be comparable, and b) a conservative upper age limit of 60 years be implemented, in particular for studies that use psycho- or neuro-linguistic tasks.

[1]Note that for the purpose of the present text the discussion is limited to healthy adult volunteers. Age at testing may become a much more complex issue for studies wishing to include child or adolescent participants and compare their linguistic performance to that of adults, since it is often a great challenge to develop tasks that are suitable across age ranges from childhood to mature speakers. Furthermore, ethical approval is often more difficult to obtain for studies that seek to investigate younger participants; and the same is true for investigations of pathological conditions, such as aphasia. These restrictions and procedures should be kept in mind and addressed at an early stage of the research design in studies intending to target populations other than healthy adults.

Third, it is important to take into account the **education levels** of participants. Since the most frequently studied population in investigations of bilingualism consists of college or university students, this factor is often considered to be unproblematic. Investigations focusing on ultimate attainment in late learners, however, often do not have this option, as such studies rely on long-term immersed bilinguals whose exposure started after puberty, and student populations do not have a sufficient length of exposure. Measuring educational levels in immigrant populations can often be problematic, as Schmid (2011) has demonstrated: educational systems in different parts of the world may not be directly comparable. Furthermore, migrant biographies are often far less straightforward than those of individuals who have lived their entire life in the same country, since migration can be a disruptive event for a variety of reasons. That notwithstanding, education may affect performance on a wide range of tasks, in particular those with a written component or those that resemble test taking to any degree, making it important that it is carefully assessed and, as far as possible, comparable across populations.

2.3.2 Intelligence, Working Memory and Other Cognitive Factors

In addition to biographical factors such as aging or education, there are also long-term differences in cognitive capacities, including intelligence and working memory (WM), which may affect (aspects of) language learning and performance. The underlying assumption in much research on bilingual populations is that experimental designs should probe participants' ability to correctly identify or use particular linguistic structures. The task should ideally measure the mastery of this structure, and nothing else. However, a number of confounding factors can have an impact on the performance on any given task, facilitating the performance of some individuals while underestimating the actual level of others, irrespective of their underlying proficiency. As Bialystok and Hakuta (1994) point out, for example, the grammaticality judgment task used by Johnson and Newport (1989) was an extremely long one, comprising 276 items, and this might have disadvantaged the older participants who might not have been able to sustain full concentration across the entire test, while younger participants presumably had less trouble with fatigue (p. 70).[2]

There is ample evidence showing that individual capacities can enhance or constrain performance on many of the tasks frequently employed in linguistic research, such as acceptability judgments. For example, differences between high-

[2]Another potential age effect unrelated to the Critical Period is connected with education: the fact that most of the younger participants in the Johnson & Newport study were college students at the time of testing. Presumably, this population has recent experience with extended testing situations from which they would benefit.

and low-WM span individuals have been found with respect to interpretation (Miyake et al. 1995) and processing (Friederici et al. 1998) of complex sentences in the L1. Grammaticality judgments and comprehension tasks can also be affected by loading working memory with concurrent memory tasks (Blackwell and Bates 1995).

Despite the desirability of controlling for background variables associated with intelligence, working memory and other cognitive factors, the tests used to assess them present a number of challenges, first among them the considerable debate over what exactly they consist of, how stable they are among individuals, and particularly how to measure them even for monolingual populations. The challenges become even more formidable when it comes to comparing monolinguals and bilinguals. It is often debated which indicator of cognitive ability/potential is the best predictor of the level of performance on linguistic tasks. The question of which measure should be chosen is not an easy one for any linguistic investigation, since WM capacity and intelligence are not only strongly correlated (Conway et al. 2003) but it is also very controversial what subcomponents make up WM. Some studies find no evidence for such a specific WM capacity (e.g. Vos et al. 2001). Additionally WM and other cognitive measures seem to vary over time (the test-retest validity of the reading span test used by Caplan and Waters 1999 is as low as 0.41), meaning that their predictive power is concomitantly lower.

These controversies notwithstanding, given the impact of individual capacities on language processing in monolinguals, it is evidently important that this factor be measured and taken into account in linguistic experiments. However, where bilingual populations are concerned, this is anything but straightforward. Measures of intelligence have long been recognized to be problematic for bilingual populations. On the one hand, administering an IQ task in a bilingual's weaker language will inevitably lead to a depressed score. This fact resulted in the assumption, held for many decades of the 20th century, that bilingualism is detrimental to intelligence (for an overview of the debate see Pavlenko 2011). This problem cannot trivially be solved by providing bilinguals with a version of the test in their stronger language for two reasons: First, questions on verbal IQ tasks often have a cultural bias. For example, Baker (1988) points out that the question 'Who discovered America?', which forms part of the Wechsler Intelligence Scale for Children, has different acceptable answers for English and American children on the one hand and Welsh children (who are taught that it was discovered in the 12th century by Madoc, Prince of Gwynedd) on the other, while neither of these answers might be acceptable to Native American children (p. 12). And, of course, there are many parts of the world where Christopher Columbus is not the household name that it is for most Western cultures. It is virtually impossible to ensure that IQ tests across languages and cultures are identical in their difficulty and validity (Baker 1988: 12). The second problem with testing bilingual participants' IQ in their strongest language is that the question 'What is your strongest language?' may vary across domains. A speaker may, for example, have higher verbal skills in one language but prefer to do math in the other. All in all, these considerations suggest that IQ tasks and any other task that is administered verbally may not be suitable to establish a

valid range of individual cognitive capacity across monolingual and bilingual participants.

The same considerations are true for measurements of verbal WM, since bilingual individuals vary strongly in their performance on WM tasks administered in the L1 versus the L2 (McDonald 2006; Gass and Lee 2011). WM capacity is usually higher in the L1, but has been shown to positively correlate in the L2 with performance on sections of the TOEFL and L2 reading comprehension abilities (for a review, see Miyake and Friedman 1998), processing of gender and number agreement in the L2 (Sagarra and Herschensohn 2010), the ability to make use of interactional feedback in L2 classroom settings (Mackey et al. 2002) and general L2 proficiency (van den Noort et al. 2006). These findings point to some kind of trade-off effect: As processing in the L2 becomes more automatized with higher levels of proficiency, WM capacity increases, while the more controlled processing necessary for less proficient learners results in less storage space, in line with Baddeley's (2003) model (see Gass and Lee 2011). Where WM is measured with respect to the verbal component, proficiency and WM capacity may therefore be confounded for the L2 populations.

Given all of these considerations, it would seem that there is no ideal way of measuring individual cognitive capacities, even though these may constrain performance on linguistic tasks in studies which aim to compare monolingual and bilingual populations. Although the comments above rule out intelligence tests in general and many of the commonly used WM tasks of verbal recall, such as the serial-recall task (Daneman and Carpenter 1980) or the reading-span task (Conway et al. 2005), there is nevertheless one possible WM option, the *n*-back digit span task (Kirchner 1958). It should be admitted that this is probably not an optimal measure of WM, but that in the context of bilingualism research it very likely is the least-worst one.

It was for this reason that the *n*-back task was included in the project on which the present volume is based. In our version of the task, the participant was presented with a sequence of digits. Each digit was displayed on the computer screen for 500 ms, followed by a white screen for a period of maximally 1500 ms. During this interval, the participant had to indicate by means of pressing a button whether the current stimulus was a match for the one presented n steps earlier in the sequence. If this was the case, the target answer was yes (pressing a green button), while a mismatch was indicated by pressing a red button. The load factor *n* was adjusted to make the task more difficult: Participants completed two blocks of 2-back trials (a total of 104 trials) and two blocks of 3-back trials (162 trials). The script for this task is available in the online supplementary material.

2.3.3 Attitude and Use

The individual cognitive factors discussed in the previous section were acknowledged to be extremely problematic for bilingualism research. Not only do these

factors vary over time, they are also difficult to measure in a way that is valid and reliable for both monolingual and bilingual populations. These difficulties notwithstanding, in this area there exist established, widely used and standardized tasks which, at least in theory, allow them to be measured. The same cannot be said for another set of factors which impact strongly on the language development and proficiency of bilinguals, namely those predictors that are connected to patterns of language use and emotions.

The literature on the role of attitudes towards L1 and L2, as well as the frequency of language use in various domains, for the development of proficiency is vast (for an overview see Schmid 2011: Chaps. 7 and 8). One of the largest problems for linguistic research is that, unlike the factors mentioned above, these predictors cannot be independently measured. Researchers have to rely on self-assessments and self-reports, which are notoriously unreliable and often affected by how participants wish to be perceived, not by how they actually feel and behave. For example, a speaker who feels (for whatever reason) that her L2 proficiency should be better than it actually is might downplay the frequency of use of that language, and a participant who wants to be polite might not be entirely honest in reporting her attitudes towards the language and speech community with which the researcher is affiliated.

Furthermore, such reports can only provide a snapshot of attitudes and habits at a specific point in time and the reliability of reports of past experiences may decrease even further. Attitudes and the relative dominance of use of the various languages are extremely fluid and changeable over time. Instructed learners, for example, often develop their attitudes based on how much they like or dislike their teachers, so that a change in the instructor may lead to a reversal of the emotional stance. Migrants may arrive in their new country filled with enthusiasm, but become disenchanted through negative experiences. Use is possibly even more variable; language habits will change as people enter a speaker's social network, or disappear from the circle of friends, causing continual fluctuations in the actual amount of use of various languages. Technological and infrastructural changes over the past decades (internet, e-books, cheap telecommunication and Skype, cheap travel) further imply that the communicative possibilities open to most migrants are completely different today from what they were even two decades ago.

Lastly, one-dimensional measures of use of L1 and L2, which ask participants to estimate the proportion of use of both languages on a typical day, cannot capture the complex interplay of language use across a variety of domains—for example interactive versus receptive (reading, watching TV), written versus auditory, formal versus informal, and so on. Whether, for example, a participant routinely uses one of her languages in the workplace or at home with her partner may impact in very different ways on her acquisition of the L2 and the attrition of her L1 (see Schmid 2007, 2011).

All of these considerations imply that any attempt to comprehensively elicit a picture of both the current status of the participants' attitudes and language use and the development of these characteristics over time would necessitate a formidably complex instrument that would be extremely time-consuming to administer (and

might still not be entirely reliable). Again, a compromise is necessary, balancing available time versus the most desirable information. These considerations have lead to a questionnaire with some 100 questions, which is available in the online supplementary material. This questionnaire takes 45–60 min to administer and will allow you to obtain a comprehensive picture of attitude and use over time, contexts and modalities.

2.4 Language Proficiency

That language proficiency impacts strongly on performance on linguistic tasks and on language processing appears to be a trivial truth. However, it is an important consideration for assessments of L2 acquisition. There are two issues for which proficiency is particularly important. The first one concerns the comparison of groups with different L1s or L2s aiming to establish whether the developmental trajectory is similar for a specific aspect of language for various populations or whether path and outcome are affected by the L1 or L2 involved. In these cases, some more **general measure of proficiency** is essential in order to determine at what stage of language acquisition individuals are and to make sure that group differences are not confounded by proficiency. As well as providing valuable information for the single study, this general measure can act as a cross-study titration for meta-study analyses.

Second, there is the question of whether there are limits to the attainability of ultimate nativeness after a certain age, in particular where specific grammatical structures are assumed to be problematic. There are two important considerations here: First, L2 proficiency and age at acquisition are almost invariably negatively correlated in populations of learners. However, a substantial proportion of the decline in eventual proficiency among older learners is clearly due to changes in learning contexts, cognitive abilities and motivation, which co-occur with age and are associated with success in language learning (Bialystok 2001). Given these explanations for decline, identifying a limit to L2 development which is caused by independent maturational processes becomes difficult. It is therefore necessary to make sure that the L2 learners investigated are not only at the top of their cohort, but also at a level of **general proficiency** that is comparable to that of the reference population(s), i.e. that the population are near-natives.

Second, recent research has established that language processing and performance on certain tasks may also vary considerably within monolingual populations, and that one important contributing factor here is also language proficiency (2015). The general assumption of homogeneity within monolingual populations may be due to the fact that participants in linguistic experiments are so often university students—that is, groups of age-matched, highly educated and relatively young individuals who can be assumed to be comparatively highly proficient speakers of their L1. If such individuals are compared with monolinguals at lower levels of proficiency, very interesting differences emerge. For example, an ERP experiment

by Pakulak and Neville (2010) finds that high- and low-proficiency monolingual speakers of English differ substantially in their brain responses to sentences which violate English phrase structure rules. The differences between high- and low-proficiency native speakers found in this study are strikingly similar to those found between natives and second-language learners in many studies. Such findings indicate that the differences found between L2 populations of different ages of acquisition, such as the behavioral results reported by Johnson and Newport (1989) or the neurolinguistic findings presented by Weber-Fox and Neville (1996), may not solely be due to AoA but to the proficiency levels in the various age groups (which decrease with AoA in the Weber-Fox and Neville study, as was pointed out by Gillon Dowens et al. 2010; van Hell and Tokowicz 2010). Similarly, Dąbrowska (2012) suggests that individual differences between speakers, such as the level of education (and, by extension, the familiarity with formal, written registers) can lead to different underlying grammars even within monolingual populations.

For studies attempting to investigate whether there are any qualitative differences between (late) L2 learners and natives with respect to certain grammatical rules or structures, it is therefore of vital importance that the proficiency level of these populations be clearly described. Tremblay (2011) has treated this issue in detail and demonstrated that the common procedure of estimating proficiency based on hours of instruction is not reliable. Furthermore, this measure is obviously not useful in determining levels of knowledge of uninstructed learners, attriters or monolinguals.

We recommend a three-tiered approach, for practical reasons. First, recruitment should encourage participants to self-select on perceived level of proficiency, that is, the recruitment text should mention that the study will test advanced or very advanced speakers. Second, we highly recommend an online screening test before participants are invited to the experiment. For many languages, standardized instruments are available, for example the Goethe-Test for German (www.goethe.de/lrn/prj/pba/deindex.htm) or the DIALANG Placement Test for Danish, Dutch, English, Finnish, French, German, Greek, Icelandic, Irish-Gaelic, Italian, Norwegian, Portuguese, Spanish and Swedish (http://www.lancaster.ac.uk/researchenterprise/dialang/about.html). Using such a test will help ensure that the participants who are eventually selected are proficient enough to take part in the study.

In order to ensure homogenous proficiency levels across populations, the actual test battery should furthermore contain at least one further measure of overall proficiency. Again, previous studies vary in which task they prefer to this end. Some have used official standardized tests for L2 learners, such as the Test of Adolescent and Adult Language (TOAL-3; Hammill et al. 1994, used by Pakulak and Neville 2010) or the Oxford Quick Placement Test (2002), which is frequently used to assess L2 English proficiency. Such tests are, however, not available for all languages. We therefore recommend the widely-used C-Test, which is easy to construct and administer (as long as the language in question has a written register) and has been shown to be a reliable indicator of proficiency at advanced levels (Schmid 2011:183). The C-Test is constructed on the basis of real examples of short

texts, which may come from different genres (newspaper articles and columns, encyclopedia entries etc.). The first sentence of each text is left intact. Starting with the second sentence, the second half of every second word is removed and replaced by a gap (compounds, proper names and words that have been gapped before are skipped and in words with uneven numbers of letters, one more letter is removed than remains standing). Participants receive one point for each word which they are able to complete correctly; an exercise which requires them to make full use of the built-in redundancy of every text and integrate their linguistic knowledge from a number of levels.

Schmid (2014) used a C-Test comprising five short texts of ca. 20 gaps each, which she administered to very advanced L2 learners of German with English as their L1 and long-term L1 German migrants in Canada. A high level of reliability was found across these texts ($\alpha > 0.8$), and for the present study we therefore opted for a shortened version, using only two texts. The texts we used to assess proficiency in Dutch, English and German can be found in the online supplementary material.

To sum up we recommend that one or more measures of general proficiency be elicited by all investigations of bilingualism for purposes of cross-linguistic comparison. In addition we suggest that if there is a standardized test for the target language, this be preferred. Lastly, we recommend that the range of proficiency in the native group be established as well, since this can clearly also vary between studies and may explain inter-study variability as much as other factors.

2.5 Conclusion

The considerations offered in this chapter with respect to external and personal factors that may influence performance on linguistic tasks underscore the complex and multifactorial nature of linguistic research. Many predictors need to be taken into account, and most of them are not easily classifiable into a few neat categories (e.g. age before vs. after puberty). It should be kept in mind that some of these factors may interact with each other in complex ways. This suggests that more in-depth insight can only be gained on the basis of cumulative evidence from a range of studies.

In order to facilitate meta-investigations of such studies, it is necessary that all predictors be stringently controlled and documented. In addition, such a multi-level approach critically relies on comparable experimental procedure, and this is a substantial challenge, in particular where data collection at different sites, in different countries and in different laboratories is concerned. This is the challenge which the following chapter will address.

References

Abutalebi, J., M. Canini, P.A. Della Rosa, L.P. Sheung, D.W. Green, and B.S. Weekes. 2014. Bilingualism protects anterior temporal lobe integrity in aging. *Neurobiology of Aging* 35(9): 2126–2133.

Baddeley, A. 2003. Working memory and language: An overview. *Journal of Communication Disorders* 36(3): 189–208.

Baker, Colin. 1988. *Key issues in bilingualism and bilingual education*. Clevedon: Multilingual Matters.

Bialystok, E. 2001. *Bilingualism in development: Language, literacy, and cognition*. Cambridge University Press, Cambridge.

Bialystok, E., and K. Hakuta. 1994. *In other words. The science and psychology of second-language acquisition*. New York: Basic Books.

Bialystok, E., F. Craik, and G. Luk. 2008. Cognitive control and lexical access in younger and older bilinguals. *Journal of Experimental Psychology. Learning, Memory, and Cognition* 34 (4): 859.

Birdsong, D., and M. Molis. 2001. On the evidence for maturational constraints in second-language acquisition. *Journal of Memory and Language* 44(2): 235–249.

Blackwell, A., and E. Bates. 1995. Inducing agrammatic profiles in normals: Evidence for the selective vulnerability of morphology under cognitive resource limitation. *Journal of Cognitive Neuroscience* 7(2): 228–257.

Caplan, D., and G.S. Waters. 1999. Verbal working memory and sentence comprehension. *Behavioral and Brain Sciences* 22(1): 77–94.

Conway, A.R., M.J. Kane, and R.W. Engle. 2003. Working memory capacity and its relation to general intelligence. *Trends in Cognitive Sciences* 7(12): 547–552.

Conway, A.R., M.J. Kane, M.F. Bunting, D.Z. Hambrick, O. Wilhelm, and R.W. Engle. 2005. Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review* 12(5): 769–786.

Dąbrowska, E. 2012. Different speakers, different grammars: Individual differences in native language attainment. *Linguistic Approaches to Bilingualism* 2(3): 219–253.

Daneman, M., and P.A. Carpenter. 1980. Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior* 19(4): 450–466.

DeKeyser, R.M. 2000. The robustness of critical period effects in second language acquisition. *Studies in second language acquisition* 22(4): 499–533.

DeKeyser, R., I. Alfi-Shabtay, and D. Ravid. 2010. Cross-linguistic evidence for the nature of age effects in second language acquisition. *Applied Psycholinguistics* 31(3): 413–438.

Dowens, M.G., M. Vergara, H.A. Barber, and M. Carreiras. 2010. Morphosyntactic processing in late second-language learners. *Journal of Cognitive Neuroscience* 22(8): 1870–1887.

Dussias, P.E. 2004. Parsing a first language like a second: The erosion of L1 parsing strategies in Spanish-English Bilinguals. *International Journal of Bilingualism* 8(3): 355–371.

Elman, J.L., E.A. Bates, M.H. Johnson, A. Karmiloff-Smith, D. Parisi, and K. Plunkett. 1996. *Rethinking innateness*. Cambridge: MIT Press.

Friederici, A.D., K. Steinhauer, A. Mecklinger, and M. Meyer. 1998. Working memory constraints on syntactic ambiguity resolution as revealed by electrical brain responses. *Biological Psychology* 47(3): 193–221.

Gass, S., and I. Lee. 2011. Working memory capacity, inhibitory control, and proficiency in a second language. In *Modeling Bilingualism: From Structure to Chaos*, ed. M.S. Schmid, and W.M. Lowie, 43–59. Amsterdam/Philadelphia: John Benjamins.

Gold, B.T., N.F. Johnson, and D.K. Powell. 2013. Lifelong bilingualism contributes to cognitive reserve against white matter integrity declines in aging. *Neuropsychologia* 51(13): 2841–2846.

Hall, K. 2008. 15 Exceptional speakers: contested and problematized gender identities. *The handbook of language and gender* 25: 353.

Hammill, D.D., V.L. Brown, S.C. Larsen, and J.L. Wiederholt. 1994. *Test of adolescent and adult language: Assessing linguistic aspects of listening, speaking, reading, and writing*. Austin, TX: Pro-Ed.

Johnson, J.S., and E.L. Newport. 1989. Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology* 21(1): 60–99.

Kirchner, W.K. 1958. Age differences in short-term retention of rapidly changing information. *Journal of Experimental Psychology* 55(4): 352.

Knecht, S., B. Dräger, M. Deppe, L. Bobe, H. Lohmann, A. Flöel, and H. Henningsen. 2000. Handedness and hemispheric language dominance in healthy humans. *Brain* 123(12): 2512–2518.

Mackey, A., J. Philip, T. Egi, A. Fujii, and T. Tatsumi. 2002. Individual differences in working memory, noticing interactional feedback and l2 development. In *Individual differences and instructed language learning*, ed. P. Robinson, 181–209. Amsterdam/Philadelphia: John Benjamins.

McDonald, J.L. 2006. Beyond the critical period: Processing-based explanations for poor grammaticality judgment performance by late second language learners. *Journal of Memory and Language* 55(3): 381–401.

McKeever, W.F., K.S. Seitz, A.J. Krutsch, and P.L. Van Eys. 1995. On language laterality in normal dextrals and sinistrals: results from the bilateral object naming latency task. *Neuropsychologia* 33(12): 1627–1635.

Miyake, A., and N.P. Friedman. 1998. Individual differences in second language proficiency: Working memory as language aptitude. In *Foreign Language Learning: Psycholinguistic studies on training and retention*, ed. A.F. Healy, and L.E. Bourne, 339–364. Mahwah: Lawrence Erlbaum.

Miyake, A., P.A. Carpenter, and M.A. Just. 1995. Reduced resources and specific impairments in normal and aphasic sentence comprehension. *Cognitive Neuropsychology* 12(6): 651–679.

Oldfield, R.C. 1971. The assessment and analysis of handedness—The Edinburgh Inventory. *Neuropsychologia* 9(1): 97–113.

Oxford Quick Placement Test. 2002. Oxford University Press, Oxford.

Pakulak, E., and H.J. Neville. 2010. Proficiency differences in syntactic processing of monolingual native speakers indexed by event-related potentials. *Journal of Cognitive Neuroscience* 22(12): 2728–2744.

Pavlenko, A. (ed.). 2011. *Thinking and speaking in two languages*. Clevedon: Multilingual Matters.

Pavlenko, A., A. Blackledge, I. Piller, and M. Teutsch-Dwyer (eds.). 2001. *Multilingualism, second language learning, and gender*. Berlin: Walter de Gruyter.

Sagarra, N., and J. Herschensohn. 2010. The role of proficiency and working memory in gender and number agreement processing in L1 and L2 Spanish. *Lingua* 120(8): 2022–2039.

Schmid, M.S. 2007. The role of L1 use for L1 attrition. In *Language attrition. Theoretical perspectives*, eds. Köpke B., M.S. Schmid, M. Keijzer and S. Dostert, 135–153. Amsterdam/Philadelphia: John Benjamins.

Schmid, M.S. 2011. *Language attrition*. Cambridge University Press, Cambridge.

Schmid, Monika S. 2014. The debate on maturational constraints in bilingual development: A perspective from first language attrition. *Language acquisition* 21(4): 386–410.

Stroop, J.R. 1935. Studies of interference in serial verbal reactions. *Journal of Experimental Psychology* 18(6): 643.

Tremblay, A. 2011. Proficiency assessment standards in second language acquisition research. *Studies in Second Language Acquisition* 33(3): 339–372.

Ullman, M.T. 2004. Contributions of memory circuits to language: The declarative/procedural model. *Cognition* 92(1): 231–270.

Van den Noort, M.W., P. Bosch, and K. Hugdahl. 2006. Foreign language proficiency and working memory capacity. *European Psychologist* 11(4): 289–296.

Van Hell, J.G., and N. Tokowicz. 2010. Event-related brain potentials and second language learning: Syntactic processing in late L2 learners at different L2 proficiency levels. *Second Language Research* 26(1): 43–74.

Vos, S.H., T.C. Gunter, H.H. Kolk, and G. Mulder. 2001. Working memory constraints on syntactic processing: An electrophysiological investigation. *Psychophysiology* 38(1): 41–63.

Weber-Fox, C., and H. Neville. 1996. Maturational constraints on functional specializations for language processing: ERP and behavioral evidence in bilingual speakers. *Cognitive Neuroscience* 8(3): 231–256.

Chapter 3
The Multi-lab, Multi-language, Multi-method Challenge

Bregtje J. Seton and Laurie A. Stowe

Abstract In this chapter, the focus will be on the general challenges that are involved in setting up, planning and running an experimental investigation of bilingual development which involves collecting data at more than one site. Some of the issues we flag up may seem logical, straightforward and obvious. However, there are obstacles that are easy to overlook when planning a multi-site study. Many of these involve tasks and decisions that take up more time than initially seems possible. In order to help you avoid unexpected problems, this chapter will sum up different issues that should be kept in mind from the start and offer a timeline for these issues which is intended to help you plan your own study and avoid unnecessary delays.

Keywords Multi-lab studies · Bilingualism · Second language acquisition · First language attrition

3.1 Introduction

First of all, it should be emphasized that in an ideal world, collecting data at multiple sites should be avoided since such an approach complicates matters considerably. However, as was pointed out in Chap. 2, the multifactorial nature of bilingual development makes it inevitable in some cases to recruit participants at different locations, in different countries and even on different continents. For example, in the language acquisition and attrition study which underlies the considerations set out in the present volume, the participants were tested in two different languages (German and Dutch) and included native speakers, immersed L2 speakers (in Germany and the Netherlands) and L1 attriters (in North America). Such a setup makes testing at a single site impossible, but allows the researcher to gain insights that would not be acquired if data from only a single linguistic context were used.

M.S. Schmid et al., *Designing Research on Bilingual Development*,
SpringerBriefs in Linguistics, DOI 10.1007/978-3-319-11529-0_3

3.2 How to Choose Partner Centers: Not All Labs Are Equal

If the decision to test in multiple centers is unavoidable, the three criteria for choosing partner centers are (a) availability of equipment needed for the study, if this cannot be transported, (b) availability of the target population, and (c) last but certainly not least, access to the lab. In our case, in order to test on-line processing, we chose to use both event-related potential (ERP) recording equipment and eye-trackers, neither of which are readily portable (in particular where air travel is involved). This also meant we usually had to make agreements with more than one investigator or lab at different testing sites. We furthermore opted to conduct our study chiefly in large urban centers with sufficiently large immigrant populations to allow us to recruit the necessary numbers of attriters and L2 speakers successfully within a reasonable period of time (which ruled out some otherwise potentially ideally suited collaborators situated in remoter areas, such as Pennsylvania State University).

Access is the criterion that leads to the most difficulties and requires the most extensive forward planning. It is important to come to an arrangement with partner centers as early as possible in the process in order to guarantee feasibility. In most cases, existing links with colleague investigators will help facilitate access. However, even labs where no such pre-existing links exist are often willing to collaborate, especially if their labs are not fully booked and you have the necessary resources and budget to pay for the use of their equipment. Since the cost of access to the lab(s) will often form a sizeable proportion of the project budget it is important to negotiate this at an early stage of planning. Furthermore, time must be available in the host lab's schedule during the periods when testing is planned. If many studies are already being carried out, this may cause scheduling difficulties. Again, early planning makes a difference.

In the study described here, four different countries were involved: Germany, the Netherlands, Canada and the US. In total, the number of labs at which testing took place was nine: Berlin,[1] Hamburg[2] and Mainz[3] in Germany; Groningen[4] and Leiden[5] in the Netherlands; London, ON[6] and Toronto[7] in Canada; and Chicago[8] and New York City[9] in the US.

---

[1]Technical University of Berlin, Dept. of Cognitive Modelling in Dynamical Man-Machine Systems.

[2]University of Hamburg, Department of Biological Psychology and Neuropsychology.

[3]Johannes Gutenberg University Mainz, Department of English and Linguistics.

[4]The Neuroimaging Center of the University of Groningen and University Medical Center.

[5]University of Leiden, Faculty of Humanities.

[6]Western University (UWO), Department of Psychology.

[7]Rotman Research Institute at Baycrest Hospital.

[8]University of Illinois at Chicago, Department of Psychology.

[9]City University New York, The Graduate Center.

3.3 Data Collection at Different Centers

When collecting data at multiple sites, it is inevitable that more than one person will be involved in testing. Even when members of a large-scale project collaborate closely it is important to agree upon a detailed protocol before testing begins. Otherwise, differences in testing method will creep in which may initially appear trivial but will later on compromise comparability of the data. The protocol should thus consist of a step-by-step description of all tasks, from the information to be given to participants up to and including which buttons participants should press in order to initiate a response and how the equipment is to be set up and calibrated in order for the experiments to run properly.

As was noted above, the key to collecting data at multiple centers is good planning. This entails that time must be allocated in the workflow for preparations so that everything is ready to go when lab space becomes available according to the agreements with the host lab. Flexibility and willingness to cooperate are essential for the guest researcher, as is sensitivity to the priorities of the partner lab: Research being carried out by an external researcher will generally, and understandably, take a back seat to experiments conducted by the host lab's own staff and students. In order to ensure that the study can be run at any time that the lab becomes available, attention should be paid to a number of aspects.

Hardware: First, the experiment should be ready to run. It is advisable, where possible, to bring equipment, such as cameras, audio recorders, and laptops for the presentation of the experiment and/or the acquisition of the data. No further preparation is then necessary and this makes it easier to use exactly the same procedures for the data processing and analysis. However, bringing equipment is not always possible, and where the experiment has to rely on the equipment available in the host lab, some level of variability is inevitable. As long as the hardware is reasonably compatible, this should not be an insurmountable issue, and Chaps. 5 and 6 below set out some of the steps that need to be taken in order to ensure that data collected at multiple sites by means of eye-tracking and EEG is comparable.

The degree of difficulty of combining data from multiple sites may also depend on the effects that a study aims to investigate. For example, our eye tracking study used a visual world paradigm with comparatively large areas of interest (AoI) on the screen. This is easier to compare across different eye trackers than for example a reading study where gaze direction is measured based on much smaller AoIs on the screen (see Chap. 5). Similarly, in our EEG experiment, we were chiefly interested in two components (the N400 and the P600) which have large amplitudes and broad scalp distributions (see Chap. 6). This made it easier to compare across hardware than if we had aimed for a component with a more limited scalp distribution and smaller amplitude where exact equivalence of the set-ups (such as the placement of the electrodes) might have been more important.

Software: It is advisable, wherever possible, that the same software be used in every lab, software being easier to transport than hardware. That said, the recording

software (which acquires the actual eye movements or the actual EEG) usually comes with the recording hardware, and we do not recommend trying to switch to a different software unless extremely good technical support is available for this purpose. Using the same presentation software, on the other hand, will not only save having to re-program experiments (which may be very labor-intensive), it also guarantees comparable presentation parameters.[10] If installation of the presentation software in the lab is not possible, the use of a laptop with the software used in the experiment as the presentation computer may be the best option. However, even if the same presentation software is available at all labs, the experiments have to be checked and adjusted for use. For example, different eye trackers need hardware-specific plug-ins for presentation software, such as E-Prime to coordinate with the eye-tracker, and different EEG systems and set-ups require specific codes in the E-prime script to ensure that the stimuli can be correctly identified in the EEG output. This means that information on matters such as port names, trigger codes or port resetting, may be necessary in order for the script to run on the system. Since timing is central for EEG measurements, the modifications will have to be double-checked before data collection can safely proceed.

Before the actual data collection begins, it is also important to have procedures established for converting the data from different labs to a single format for later analysis. It is advisable to run a pilot study and compare the data between centers in order to ensure that the presentation software, its interface with local recording equipment and the data conversion all work properly. This will prevent unpleasant surprises, in the worst case after data collection has finished and adjustments are no longer possible.

In order to resolve these problems, it can be advisable to hire a local research assistant who is familiar with the lab and the setup. In such cases, it should be kept in mind that an extra induction period may be necessary to train the RA on the use of any specific software and the experiment itself, and for the project team member to be trained on the use of the local hardware. Even researchers familiar with one type of eye-tracker or EEG recording system may require training in order to be able to use another. Using other people's set-ups requires extra care and therefore extra training and sometimes additional *insurance* may also be necessary. Here it is important to bear in mind that some insurance companies in other parts of the world will not automatically cover damages incurred in North America due to the high claims culture in the United States and Canada. Such considerations may seem like minor details, but if not planned for, they may cause substantial delays and additional expense.

Data storage and back-up: The data collected from multiple sites will need to be stored and backed up in one place. Multiple project members may need to have access to the data, which makes it necessary to have a well-organized, structured and secure back-up system. With data being transferred from one institute to the

[10]In the present study, E-Prime (1 or 2, Psychology Software Tools) was used to program all experiments.

other, it is crucial to keep in mind that data have to be anonymized so that participants' personal information cannot be connected with their names. Secure data storage and encrypted transfer are non-negotiable when it comes to any data where individuals are concerned.

3.4 The Role of Local Assistants

Local assistants can make a substantial difference to the ease with which a multi-lab study is carried out, in that they not only reduce the expense incurred by the project member's stay but can also decrease the duration of the fieldwork by a wide margin. However, since assistants who are hired locally are not as closely involved in the project's genesis, a clear, detailed and fail-safe protocol becomes even more essential. The ideal candidate for such a research assistantship is someone who is already trained in the lab. Hiring assistants from the host institution, who can also help familiarize the project member her- or himself with the equipment and protocols of the local lab if necessary, will often save time. It should be kept in mind, however, that the assistant does not have final responsibility for the data and the project member responsible for testing on location must acquire the necessary expertise before data acquisition commences.

A local assistant can also facilitate recruitment of participants well in advance. Recruitment is often a time-consuming procedure, in particular where very specific requirements obtain regarding the testing population, so the earlier it can be started, the better. Local assistants not only often have better access to the target community, they can also be indispensable as speakers of the target languages in cases where the project member does not have the necessary proficiency. In the example study, the project members responsible for data collection in North America, where both Dutch and German attriters were tested, could not be native speakers of both languages, and assistants who complemented the native speaker criteria were a valuable addition to the team.

These requirements imply that it may often be necessary to find more than one assistant, in order to provide the necessary skills, both technical and linguistic, in particular since their availability for the entire duration of the project may be a further issue. Additionally, the host institution may have specific requirements research assistants need to comply with before they can start working at the site, for example, following a required training, having certain immunizations (TB, Hepatitis B) or being screened for any criminal records. Lastly, native speakers of some languages may be precluded from taking on paid employment due to visa restrictions. Taken together, these concerns can make finding and training assistants time-consuming, but suitable assistants may well be crucial to the project's success.

3.5 Planning Ahead: Visas, Ethics and Hurricanes

There are several more factors which can interfere with planning during the preparations before going to the host institution. Where multi-country studies are concerned, it is important to be aware of and plan for visa requirements and to assess these at an early point so as to allow sufficient time for any complications that may arise, since unexpectedly long visa procedures can easily delay the onset of the study. Another potentially serious complication for multi-lab settings is formed by experimental ethics procedures, particularly if some of the labs are in different countries where the standard procedures are less familiar to the research group. In North America, for example, the ethics boards follow national guidelines which often require completing an online course before an ethics application can be approved (information about the relevant requirements can be found on-line or through contact with the host lab).

Furthermore, ethics approval procedures tend to differ between pure research institutions and medical facilities. EEG labs are frequently part of an institute situated in a hospital environment and therefore fall under the medical ethics board, whose requirements are typically more stringent due to the potential involvement of patients. The ethics approval procedures may take a substantial time (in our experience up to six months) before permission to test is given, but it is crucial that they be scrupulously followed and applied. This also means that participants must (a) be fully informed about the experimental procedures beforehand, including the discomforts of EEG measurements and that recordings will be made, (b) realize that they are not bound in any way to complete the experiment if they choose not to, and (c) be aware that care will be taken with any personal data that they share so that it cannot become public.

Despite all we can do to prepare, the actual duration of the testing period will almost certainly be longer than originally planned. It may be more difficult than anticipated to recruit the target population and therefore necessary to continue testing longer than originally intended. The host institution should always be appraised of any such potential delay. Furthermore, it is wise never to discount the possibility of other events that are outside of the control of the researcher. It may seem far-fetched to anticipate, for example, the impact of a natural disaster such as a hurricane, a flood or an earthquake, but as we found out within our own study, if one does take place it can seriously throw out the project schedule. Obviously, participants in areas affected by such events may be unavailable for months to come. So, even when devising the best laid plans, keep in mind that natural disasters, illness, and many other things may potentially interfere to an unpredictable extent.

3.6 Checklists[11]

Preparation Steps	
Contacting possible host labs for accessibility • time periods available • contacts for further technical information	☐
Drawing up experimental protocol • Detailed protocol for all aspects of data collection and storage, with site-specific annotations • Secure storage protocol with clearly defined structure for remote storage of data • Protocol for anonymization of participant information	☐
Determine what set-up is used in the lab (where relevant): • how compatible is the hardware? • is the same presentation software being used at all sites? If not, can presentation software be installed? How does presentation software communicate with hardware and recording software? • how easy is data conversion for joint analysis?	☐
Estimate time necessary on site in order to • learn different set-up, adapt experimental presentation • train assistants and adapt protocol with local details • recruit and run participants: if population is limited, more time will be necessary	☐
Check visa regulations for working in the lab (where relevant) • estimate time necessary to fulfill requirements • remember multiple-entry visa may be necessary	☐
Check ethics procedures and other requirements of the host institution: • estimate time necessary • check what the restrictions are for studies (e.g. certain questions in questionnaires or certain tests or time limits) • check what the restrictions are for data storage and transfer	☐
Check procedures for hiring research assistants • estimate time necessary	☐

[11] Please note that these steps are not necessarily in chronological order. Each project will probably need to revise them at various stages, as they are often dependent on one another.

Budget checklist

Project member's costs • **e.g. travel, accommodation, car hire**	☐
Buying or paying for equipment (where relevant) • **e.g. caps for EEG**	☐
Research assistant (where relevant)	☐
Expenses of lab use • **e.g. price per session, costs of consumables**	☐
Check whether insurance is necessary for use of the lab of for participants	☐
Participant expenses • **e.g. reimbursement, travel expenses, parking**	☐

Chapter 4
Collecting and Analyzing Spontaneous Speech Data

Christopher Bergmann

Abstract Spontaneous speech samples can serve several purposes within a language acquisition study. First, spontaneous speech can be used in various ways to assess the proficiency of individuals, for example via accent ratings, measures of lexical variability, and frequency of different types of errors. It can also be an object of analysis in its own right, for example in examining interactions between linguistic variables and potential trade-off effects, including factors such as the use of fixed expressions, sentence complexity or speech rate as well as the interaction between them. Finally, it can provide a comparison of production with processing or comprehension measures. While the analysis of this sort of data is not, in principle, different for a large scale study like this than for a single dedicated study, the versatility of information it allows the researcher to assess is easy to underestimate. This chapter will discuss the collection, transcription, coding and analysis of spontaneous speech samples, serving as an exploration and reminder of some of those possibilities as well as an introduction to how to make use of them.

Keywords Bilingualism · Methodology · Free speech · Grammatical gender

4.1 Introduction

Free speech can form a valuable addition not only to large-scale studies, such as the one described here, but also to dedicated studies which do not focus on spontaneous data in any of the ways listed above. As we stressed in Chap. 1, dedicated studies of microlinguistic features can be difficult to integrate into the literature as a whole because they differ on too many variables. It may therefore prove fruitful to have free speech data available, since such data may be used to calibrate the speaker characteristics of particular investigations for a broader comparison or meta-study.

The first section of this chapter will discuss a number of linguistic phenomena that can readily be investigated in free speech data as well as some of its limitations.

M.S. Schmid et al., *Designing Research on Bilingual Development*,
SpringerBriefs in Linguistics, DOI 10.1007/978-3-319-11529-0_4

In the second, we will address practical matters related to the collection, preparation, transcription and analysis of free speech data. For all these procedures, in particular where the transcription format is concerned, a range of options exist. We will limit our discussion to the CHAT system—by no means the only one available, but one which is widely used, well documented, flexible and user-friendly. Many of the suggestions we make here apply equally or similarly to other formats. The final section addresses how the data can be annotated for actual analysis, exemplified by a coding system that we have devised for coding gender errors. The term 'coding' is used to refer to manually adding a layer of information to transcribed data, in this case morphological information about grammatical gender in German speech samples.

4.2 Areas of Investigation

As was pointed out above, free speech lends itself to analysis on a wide range of linguistic levels. Since an exhaustive discussion of these matters is beyond the scope of the present text, we will exemplify applications based on three areas which typically play a large role in language acquisition research: phonetics, speech disfluencies, and lexical choice. We will highlight some aspects of those phenomena that can most readily be investigated using free speech data.

4.2.1 Phonetics and Phonology

Recordings of free speech, provided they are of sufficient quality (see below), are very well suited for investigations of phonological and phonetic questions. The main advantage of this type of data lies in the fact that it investigates the naturalistic production of speech sounds, a context that is less open to the application of explicit knowledge and conscious monitoring than many other elicitation tasks (e.g. word list reading), which foster slow and careful enunciation, artificially producing or magnifying phonetic and phonemic distinctions that are obscured or absent in natural speech. Free speech also elicits phenomena that are idiosyncratic to fast, connected speech.

One disadvantage of using this type of data lies in the fact that it is less easy to ensure the actual production of the intended phonemes or phones. This can lead to problems in cases where the elicited speech samples do not contain a sufficient number of the target speech sounds or of the various contexts in which they occur. In particular where less frequent sounds or sound combinations are concerned, it is wise to assess the reliability of their occurrence based on pilot recordings and ascertain that each speech sample will contain at least twice the number of productions that are deemed necessary for analysis. If this turns out not to be the case, further methods of elicitation should be considered (see below).

A further advantage of free speech is that, unlike data from more targeted elicitation methods (such as list reading or grammaticality judgments), it can be used as the basis for global native speaker ratings of naturalness or nativeness/accentedness of pronunciation. Such ratings can then be compared to the acoustic measurements.

4.2.2 Disfluencies

Fluency is a second area of investigation which has been much addressed, in particular in bilingualism research, over the past decades, and for which recordings of free speech are crucial. It is an inherent characteristic of free speech that speakers —including monolingual natives—hesitate, pause, stumble over their words, start over and derail their own trains of thought much more often than in other types of task (e.g. reading aloud tasks). These disfluencies have captured the interest of a large psycholinguistic community and are considered a window on both the mental representation of language and the organization of speech production processes.

Disfluencies are ubiquitous in spontaneous language use; they have been estimated to affect 5–10 % of all words and up to one third of all utterances (Shriberg 2002). They can be coded in a fairly straightforward way once the recordings have been transcribed (see below). What makes disfluencies so interesting is that, at first sight, they only seem to be random interruptions in the speech stream. In fact, they have been shown to be intimately linked to the language-specific structures that are being produced. As a consequence, to the extent that second-language learners are not native-like, they do not hesitate and stumble in the same way as a native speaker would (Hieke 1981).

4.2.3 Lexical Variability

A third area of investigation in which free speech data have been widely used concerns lexical choice. Range and accessibility in real time of the mental lexicon vary substantially between populations, and free speech data make it possible to estimate these characteristics. Again, it is an advantage that spontaneous recordings of more than a few minutes of naturalistic speech typically contain a sufficient amount of data, even without special elicitation techniques, and that these data can be extracted conveniently once the recordings have been transcribed. A very common measure is lexical diversity, that is, the range of the productive lexicon that is being used. High lexical diversity is associated with a sophisticated vocabulary, the use of low-frequency items and a low rate of repetition, while low-diversity speech samples contain only more frequent words with many repetitions. Lexical diversity has been found to be reduced even in highly proficient second-language learners as opposed to native speakers (Sanz Espinar 1999; Noyau

and Paprocka 2000). This aspect can be particularly interesting for research that is based on longitudinal acquisition data or the speech production of impaired populations. Using software like CLAN (see below), both the traditional type/token ratio (i.e. the quotient of distinct words and the individual occurrences of these words in a sample) as well as more sophisticated estimates of lexical diversity, such as D (for a comparison of some methods see McCarthy and Jarvis 2010), can be calculated automatically. This software can also handle the necessary preprocessing steps, such as lemmatization (so that different inflected forms of the same lemma are not counted as two different words) for a number of languages.

Free speech samples are also useful for investigating lexical access at the level of phrasal chunks (Ellis 2003), that is, of lexical elements that form larger units. Just like disfluencies, such phrasemes—or fixed expressions—are of interest for psycholinguists. It has been assumed that spontaneous speech production is speeded up by the ability to retrieve not only individual lemmas from the mental lexicon, but also groups of words that contribute to a shared meaning. Since these phrases have a 'precompiled' structure and meaning, a higher density of these phrasemes might be correlated with an increased articulation rate; this can easily be tested using recordings of free speech. However, it has to be taken into consideration that coding fixed expressions in free speech cannot be done automatically and is therefore more effortful and time-consuming than simple probes into lexical diversity, which can be fully automated. A clear definition of what constitutes a fixed expression or chunk is also necessary (for an overview, see Wray 1999, 2002).

4.3 Elicitation and Data Collection

4.3.1 The Film Retelling Task

In the study that this book is based on, we made use of a film-retelling task that has been used frequently in second-language research. The paradigm we used was developed in the 1980s as part of a European Science Foundation project on L2 acquisition among adult immigrants (Klein and Perdue 1989). It consists of showing participants a 10 min extract from a film and audio-taping them while they retell the content of the sequence immediately afterwards. The excerpt originates from the silent film 'Modern Times' (1936), starring Charlie Chaplin and Paulette Goddard; it starts around 33 min into the film. There are a number of advantages associated with using this particular film and sequence: First, data elicited in this manner is comparable to a large and growing number of datasets in a variety of linguistic fields; second, the fragment has a balanced distribution of action-packed scenes on the one hand and more reflective episodes on the other, giving the participants the option both to describe the actual storyline and to muse about the significance of some more symbolic parts of the film, eliciting a wider range of lexical items; third, 'Modern Times' is a silent film, making it accessible to a wide range of audiences and offering

little to no support for lexical access. It does contain some music as well as a small number of English-language intertitles, but neither are crucial for understanding the plot of the sequence, which can be summarized as follows:

> Charlie Chaplin, newly released from prison, takes a job at a shipyard and, while trying to find a wedge-shaped piece of wood for his supervisor, causes an unfinished ship to sink. He leaves the job (presumably to pre-empt being fired) and takes a walk into town where he meets a homeless young woman who has been caught stealing a loaf of bread. Charlie takes the theft upon himself—maybe because he wants to go back to prison, maybe out of his liking and pity for the girl. The plan fails: He is set free, the girl is arrested. He goes for lunch and has himself arrested as well for not paying the bill. Using the fact that his arresting officer is momentarily distracted, he compounds his offense by stealing a cigar and some chocolate for two passing children from a kiosk. He ends up in a crowded paddy wagon, where the bread-stealing woman joins him and he re-introduces himself to her. An accident offers them the chance to escape together. Eventually, they rest on the curbside in a rich part of town, where they witness a scene of domestic bliss between a, presumably newlywed, couple. There, they fantasize about having a home of their own and becoming husband and (house)wife—a dream that is disrupted by the sudden appearance of a policeman who does not know that they are on the run, but chases them from their resting place as loiterers.

Depending on the population being tested, it is useful to consider whether the audience will be familiar with some settings or items in the film, such as a shipyard or a telephone. The film may also not be entirely suitable for eliciting free speech in children below elementary school age.

One legal caveat: In most countries, 'Modern Times' is not considered to be in the public domain and can therefore not be lawfully obtained from the internet for free; it is, however, purchasable from all major online providers of books, DVDs and film downloads. We also advise checking whether showing experimental participants an excerpt from the film is seen as a public screening under local laws; if so, it may be necessary to obtain an official authorization or pay license fees.

It is also necessary to ensure that the presentation of the sequence and the elicitation procedure are the same across all participants. We have opted for showing the fragment only once in one go, and not allowing participants to take notes while they watch. Remaining in the same room as the participants allows interaction with them, if necessary. Interruptions or interference, such as ambient noise, should be avoided wherever possible. Minor background sounds can be rendered less disruptive by using circumaural headphones (entirely enclosing the ears) instead of loudspeakers if noise in the testing area is likely. If the presentation is disturbed by any outside event or by the participants' unexpectedly diverting their attention from the film, playback should be stopped as quickly as possible and the file be rewound by about 10–15 s before it is started again. After the end of the film, participants may require a few instants to reflect upon what they saw before starting their narrative; this can occur while putting the recording equipment into position if necessary. There should, however, not be a prolonged pause after the screening, so that participants can retell what they saw fresh from memory.

During the retelling, quiet and noise-free surroundings are even more imperative to ensure good audio quality. Lab settings might seem sterile and impersonal, but in

most households and al fresco environments, it will be difficult to locate a spot where the participant's retelling is the only sound in the environment. Even a washing machine in an adjacent room, clattering dishes or bird song through an open window may cause significant data loss—and be extremely irritating for the transcriber.

While the present section focuses on the use and analysis of audio recordings, it should also be pointed out that there is considerable recent interest in speech gesture (e.g., Gullberg and McCafferty 2008). Free speech of the type that is elicited by this task is suitable for gesture analysis, so it may be advisable to not only audio- but also videotape the retelling. When videotaping, participants should be placed in front of a light, plain background, without a table in front of them, and the experimenter should be seated next to the camera, so that the participant looks toward it as a natural consequence of orienting toward the listener. Participants should also be seated in a chair without armrests, as these can make them gesture-lazy. Nothing should be within reach that could influence or interrupt their gestural behavior, such as drinking glasses, pens or other small objects that can be used to fidget with. If only an audio recording is made, comfort is the only criterion for seating.

We have discussed the importance of informed consent before (see Sect. 3.5), but we would like to emphasize particularly that participants have to be made explicitly aware that they will be audio- or videotaped. Since video recording is more sensitive for many participants, it is possible to give subjects the possibility to opt out of the videotaping and only retell the film in front of an audio recorder. If video data are more central to the whole study, this will not be an option, of course.

The retelling should put the participant at center stage; the experimenter's contribution should be kept to an absolute minimum. After making sure that all recording devices are running, a few carefully chosen standard cues to remind participants of what they are supposed to do and to prompt them to start their retelling should be employed. A typical cue would be: 'I want you to recount the plot of the film fragment you just saw. Please try to provide as much detail as you can. I will not ask any more questions once you have started talking, so just keep going until you are done. Is everything clear to you? Great, tell me then: What was the sequence about?'

Even if it seems rude, it is important to refrain from verbally interacting with the participant (and from gesturing, when videotaping the participant's gestures) to avoid interfering with the retelling. Subjects regularly address the experimenter directly or even ask for help finding words, especially in populations with less than fully native proficiency. Smile politely and try not to say a word. Most participants do not actually need the help, but will bootstrap themselves sooner or later by finding synonyms or resorting to different constructions. Only in the most extreme cases of participants getting caught in their own linguistic web, should the researcher say something like 'Don't worry, just tell me the next thing you remember.'

Sometimes participants get stalled due to being unable to remember the next scene they want to describe. Again, it is preferable to give them some time to figure

it out themselves. If it becomes absolutely necessary to put them back on track, general questions like 'What does Charlie do next?' or 'Do you remember what the girl did?' should precede actually giving them a lead on the content. If someone makes factual errors in retelling the story or mixes up the sequence of events, there is no need to point it out to them, since the content is not central to the elicitation. If important scenes are omitted, it is possible to revisit these after participants have finished their retellings, starting with unspecific questions ('I think you have not told me about the scene where Charlie does something terribly stupid, have you?') before trying to jog their memory more explicitly.

4.3.2 Ensuring Adequate Audio Recording Quality

While the recording environment is important for audio quality, the equipment is no less so. Mobile phones, headsets and other devices that can record audio or video, but are not specifically designed to do so, are not up to the task. The recorder should have two main characteristics: First, it should use a solid-state disk (SSD) instead of any other storage device. SSDs are more robust than electromechanical disks, they produce less noise while saving data (which interferes with the recording quality) and they have shorter access times. This makes them more expensive than other storage media by a factor of up to ten, but it is worth the investment, particularly since audio data do not take up a lot of storage space, so not many will be required.

Second, the recorder should support the connection of a dynamic microphone using an XLR connector. Among professional microphones, dynamic ones are most resistant to adverse conditions, such as moisture and rough treatment (and certainly more so than condenser microphones). They are also fairly affordable and do not require an additional preamplifier. XLR does not have any immediate advantage above other types of audio transmission, but it is standard for the best microphones. We used a TASCAM DR-100 recorder (in the \$200–300 price range as of 2015) as well as various Sennheiser microphones (in the \$100–150 price range as of 2015) that satisfy all the requirements mentioned above.

Professional recorders provide a choice among different file formats for storing recordings. The best choice is an uncompressed file, typically in the Waveform Audio File Format (WAVE or WAV). The file size will be fairly large (around 15 MB for 3 min of speech), but phonetic detail will not be compromised by lossy data compression. If the files are to be used for phonetic analysis, no audio compression of any type should be used. Even if this is not the original intention, free speech data can be employed for more than the original goal; recording with future applications in mind can save effort in gathering more data.

For transcription or other purposes, the files can be converted to the MP3 format or another format which takes up less space. MP3s use various compression algorithms, resulting in considerably smaller files, the size of which depends on the amount of data that is stored per second of the recording. This amount is typically expressed as a so-called bit rate in kilobytes per second (kb/s or kbps). The same

3 min file that takes up 15 MB of disk space as a WAV file can become as small as 2 MB when compressed and saved at a bit rate of around 96 kbps (the absolute minimum for retaining sufficient data). Most file converters support the use of variable bit rates for which the amount of data stored will also depend on the amount of data available in the original file. There are various other types of compressed audio files, but the advantage of MP3s is that they can be played on virtually every device that supports audio files. Again, even if phonetic analyses are not a consideration, it is best not to delete the original WAV files. This is because an accumulation of data loss inevitably occurs when processing and resaving even in highest-quality MP3s.

4.3.3 *Eliciting Specific Data*

Free speech has many strengths, but one weakness is that it may not sample enough of one specific variable of interest. We therefore briefly turn to the question of how to elicit free speech which includes more specific data, such as particular phonemes. We have had the experience that even in fairly long retellings (10–15 min), it may be impossible to find a sufficient number of analyzable tokens of some medium-frequency phonemes of German, such as /œ/, /oː/and /yː/. Having participants read out lists of sentences has the disadvantage that they are likely to be more careful, and may use conscious strategies that do not represent their standard use. The middle ground between standard retellings and lists of words or sentences would be the retelling of a more directed stimulus created for the specific purpose. Picture stories, for instance, have been found to work well. If the target is, for example, to elicit the German phonemes /œ/, /oː/and /yː/ mentioned above, one might construct a cartoon strip that revolves around two squirrels (*Eichhörnchen*) who go to the flea market (*Flohmarkt*) to sell a pile of books (*Bücher*). Such manipulations are both inconspicuous and effective. For other ideas on how to elicit particular grammatical constructions, tensed forms or referring expressions like pronouns, any number of possible paradigms ranging from relatively free to relatively fixed can be found in the literature. In particular, investigations of child language acquisition are often extremely imaginative when it comes to inventing good elicitation strategies.

4.4 Transcription

The next step is to decide on an appropriate system of transcription for the audio recordings. CHAT, the transcription system that we recommend and describe here, has a number of advantages, in particular when it comes to research on bilingualism: One of them is that files created following the CHAT format can be processed by means of the CLAN program. This free software supports automatic

assessment of the frequency of lexical items in the transcriptions, helps with morphological tagging and offers a number of other analysis options.

CHAT and CLAN were developed within the CHILDES project under the direction of Brian MacWhinney (MacWhinney 2014a, b). The system, the documentation and the publications are freely available online (http://childes.psy.cmu.edu/) and, along with our own experience, form the basis of this section. For those with no familiarity with transcribing free speech data, this crash course is meant to provide some clarity on whether delving into the voluminous handbooks is worth the effort. We cover only the basics of transcribing and do not go into the wider scope and more sophisticated applications of CHAT and CLAN; for that, the reader is referred to the publications on the CHILDES website.

The first decision to be taken is whether to transcribe orthographically or phonetically. IPA characters present difficulties for many programs, so phonetic transcriptions more usually use (X-)SAMPA or another ASCII-based transcription system. CHAT deals with any system as long as the files are encoded in UTF-8 format. This is specified in the first of the headers that mark the beginning and end of every valid file containing transcriptions in the CHAT format:

1. @UTF8
   @Begin
   @Languages: deu, eng
   @Participants: PTC FT374 Subject, INT John_Doe Investigator
   @ID: language|corpus|code|age|sex|group|SES|role|education|custom|
   …
   …
   …
   @End

The second header simply marks the beginning of the file. The corresponding header @End has to appear on the very last line of the transcript. In the third header, @Languages, all languages are indicated that are spoken in the recording being transcribed. CHAT expects language codes from the ISO 639-3 standard (www-01.sil.org/iso639-3/codes.asp). The fourth header requires three-letter abbreviations for all speakers in the recording as well as their full names using an underscore to separate the first and the last name, or (as here) participant codes and a specification of their role in the experimental setting (a list of possible roles, such as Child, Sibling, Student, Adult, etc. can be found in the CHAT manual). The fifth header is the most complex one, with ten different slots to register the corpus that the recording at hand belongs to, the sex, age and educational level of the participant etc. The first three fields—language (e.g. 'deu, eng'), corpus ('L2 speakers'), speaker code ('PTC')—and the second before last—speaker role ('Subject')—are mandatory. All other fields can be left empty, if they are not to be considered in the investigation. In headers 3–5, a tab stop, rather than a space, has to be inserted after the colon.

Everything that has been spoken and is transcribed orthographically appears on the so-called main tier or utterance tier. Every line of this type has to conform to a

basic structure: an asterisk, followed by the three letter code of the person speaking (as specified in the fourth header above), a colon and a tab stop, followed by the transcription of a single utterance (one sentence or phrase). Lines typically end with a full stop, a question mark or exclamation mark:

2. *XYZ: the transcribed text goes here.

Full stops, question or exclamation marks cannot be used within words and utterances, that is, no utterance tier may contain more than one sentence. Commas, semicolons, square brackets ([]) and angle brackets (<>) are also not accepted within words because they serve internal syntactic functions. Capital letters should only be used when the orthography for the word class in the language you are transcribing requires them (e.g. not to mark the beginning of a line or sentence). This is important if you are considering automatized morphological analysis: Regardless of the word form, any word beginning with a capital will be classified as a proper noun, even if it is, for example, a preposition beginning a sentence.

Following general orthographic rules is a good foundation for transcription when phonetic transcription is not used. Use spaces to separate words, even if the language being transcribed does not typically use them. Start a new line for every new clause or complete utterance which is less than a clause. A possible guideline is that every finite verb, along with whatever it directly governs, goes on one line. Subordinate clauses are also governed by finite verbs in the superordinate main clause, but the finite verb in the embedded clause takes precedence over the one in the matrix clause:

3. *XYZ: she wanted to know +/.
   *XYZ: +, why they were convinced +/.
   *XYZ: +, he was involved in the larceny +/.
   *XYZ: +, that was discovered by police recently.

Note that lines which contain an entire clause, but not an entire sentence end with ‘+/.’ instead of a full stop only, and that continuations within a single sentence begin with ‘+,’ to indicate which lines belong together.

One more convention to be aware of concerns the transcription of abbreviations and numbers. The recommended format—abbreviations separated by underscores and numbers written out—is seen here:

4. *XYZ: we have been living in the U_S_A for forty years.

In languages like English, compounds that are typically written as one word (such as ‘rainbow’ or ‘eyebrow’) can be transcribed following general orthography. For hyphenated compounds and compounds in other languages, a plus sign (+) is used, as well as replacing hyphens in English.

5. *XYZ: she was wearing an eccentric blue+green dress.

Names of organizations (*World Health Organization*), titles of works (*A Portrait of the Artist as a Young Man*) and similar compound phrases (*Little Red Riding Hood*)

that would normally be italicized or enclosed in quotation marks, should also be joined by underscores:

6. *XYZ: he sat in his armchair, rereading Against_the_Day.

In any language, transcribe what is actually said rather than correcting it. If a participant says 'we gonna', the transcript should as well, rather than 'we are going to'. For reduced forms, brackets serve to indicate the full form, such as in the following sentence, as an aide to interpretation at a later stage:

7. *XYZ: he (i)s convinced that we can (no)t help (th)em.

If the reduced form only bears a passing resemblance to the full form, the latter can be added between square brackets:

8. *XYZ: whadya [: what did you] do to get out there?

Some pronunciation basics can also be specified on the main tier line. This holds for stress, sound lengthening and word-internal pauses: Stress is indicated by two symbols from the International Phonetic alphabet (ˈ and ˌ), lengthening by a colon (:) and word-internal pauses by a circumflex (^):

9. *XYZ: it was an ˈabsoˌlutely hu:ge concate^nation of problems.

As we already mentioned, free speech data are typically not produced in a fully fluent fashion. This makes it important to consistently code markers of disfluency, such as ***filled pauses*** (10), ***empty pauses*** (11), ***stuttering*** (12), ***repetitions*** (13), ***retracings*** (14) and ***reformulations*** (15):

10. *XYZ: this is the place where we ah@fp worked back then.

The code 'ah@fp' corresponds to any type of filled pause (henceforth 'fp') in the recording, such as 'uhm' or 'hm'. Unless the phonetic form of the various pauses is of interest, one general label (such as 'ah') can be used for all filled pauses, followed by '@fp'. Empty pauses are indicated in the transcript by a full stop in round brackets:

11. *XYZ: I have tasted it and it was not exactly (.) delicious.

You can also put the exact measured length of the pause in seconds inside the brackets, that is, '(3.4)' instead of '(.)'.This information is much easier to retrieve during transcription than later, so considering whether to do this beforehand is advisable.

If a speaker stutters before (or even without) uttering an entire word, all failed attempts at producing the word are marked with an ampersand:

12. *XYZ: she could &re &re relate to what I had told her.

Repetitions are distinguished from stuttering, as are retracings (new attempts at the same sentence structure) and reformulations (entirely discarding the previous sentence and starting afresh). The part of the utterance that is being repeated or replaced is enclosed in angle brackets, followed by one, two or three forward

slashes in square brackets for repetitions, retracings and reformulations respectively:

13. *XYZ: he was asked <to report> [/] to report to the manager.
14. *XYZ: they <have refused to> [//] were not inclined to do it.
15. *XYZ: <I wanted to> [///] she tried to avoid this encounter.

CHAT also provides transcription conventions for recordings containing code-switching, that is, utterances in one language that are interspersed with words from other languages or varieties (16), for unintelligible speech (17), for omitted words (18) and for errors (19).

16. *XYZ: Austria is famous for its delicious Germknödel@s:deu.

Code-switches are tagged using a so-called 'special form marker', a code where an "at" sign (@) is followed by either a predefined or a user-defined code. The letter 's' in the code used here stands for 'second-language form'; as you see above, it can be followed by a colon and the ISO code of the intruding language. Other predefined codes include markers for dialect forms (@d), neologisms (@n), onomatopoeia (@o) and singing (@si).

Unintelligible forms are replaced by 'xxx' on the main tier; it is possible to add another tier (another line of information following each main tier line) to add phonetic transcriptions of utterances that do not have an obvious orthographical equivalent (see next section). On this tier, unintelligible forms are coded as 'yyy' instead of 'xxx'.

17. *XYZ: he couldn't help xxx when he heard the news.

Words that are apparently omitted by the speaker can be added to the transcript. The best available guess at which word is missing should be preceded by the digit zero (0).

18. *XYZ: she came over and prodded me in 0the back.

If the speaker produces errors, an asterisk in square brackets is added after the incorrect form. It is not strictly necessary to enclose the speech material that contains the error in angle brackets, but we recommend doing so consistently, particularly when the error is not limited to the single word preceding the bracketed asterisk, and to expedite providing a correction. Even on the main tier, more specific information can be provided about the nature of the error by using [*p] for phonetic and phonological errors, [*m] for morphological errors, and [*s] for semantic errors, which will speed searches for certain types of errors:

19. *XYZ: their performance <blowed [:: blew]> [* m] my mind!

We have concentrated on error coding, but CHAT offers a wide range of possibilities for increasing the level of detail of the transcription. The full manual provides a lot of additional information beyond the general transcription guidelines that we have outlined here. We will cover one additional possibility in the next section: how to add a tier to code specific information of interest.

4.5 Specific Annotation for Target Analyses: Gender Coding

Adding more information to a transcript can easily be accomplished by inserting a dependent tier. This is another line that appears below the main tier and can contain any kind of data desired. Unlike the main tier, dependent tiers begin with a percent sign (%), followed by a three-letter code identifying the type of tier, a colon and a tab stop. You can add as many dependent tier lines as deemed necessary to each individual utterance (but no more than one tier of the same type). There are a few standard dependent tiers, such as the %com tier for comments, the %err tier for noting more information about errors that you have already marked on the main tier, the %mor tier for morphological coding, the %pho tier for a phonetic transcription of (parts of) the utterance on the main tier and the %tim tier for timing information to keep track of the point in time in the recording that a line in the transcript refers to (very useful if it becomes necessary to check the original recording).

The example here taken from our own study illustrates a coding system that we devised for the analysis of the frequency and accuracy with which gender-marked forms are used in spontaneous speech. Grammatical gender is as pervasive as it is elusive for learners of German and Dutch, posing a considerable challenge even to the most advanced L2 speakers of these languages (see Chap. 1). We therefore wanted to know whether their productive performance in free speech distinguishes them from a native control group or an attriter group, and to what extent the mastery of productive gender correlates with other measurements of gender processing collected from the same speakers. We therefore decided to code every (pro)nominal element in the transcriptions of the retellings on a separate coding tier (which we somewhat misleadingly called %fnp for 'full noun phrase', even though pronouns were also coded here).

For reasons of brevity and clarity, the coding system, as exemplified in the present section, only gives the relevant German categories and word forms. Dutch makes fewer distinctions with respect to gender than German. For that reason, adapting the coding system to Dutch mainly involved deleting those categories that are not marked in this language.

When setting up a coding system, it is important to first compile all information that should be coded, since omissions and oversights at this stage will lead to problems later on. In our case, we were interested in (a) the grammatical form of the element to be coded, (b) the type of element, (c) its basic morphosyntactic structure, (d) its gender, number, case and definiteness (all of which relate to determining whether it is an error), (e) whether an error has occurred, and finally (f) the syntactic antecedents for pronouns.

The orthographic form is the relevant part of the utterance tier, repeated between quotation marks in the coding tier to identify the coding's referent. The type of the element can take three different levels: full noun, bare noun and pronoun. Everything that is preceded by either a determiner or an adjective is considered to be a full noun, whereas noun phrases that consist of nothing but the noun itself are

counted as 'bare nouns'. The distinction between full nouns and bare nouns is motivated by the fact that grammatical gender in German is not directly marked on the nouns themselves, but on accompanying determiners or adjectives. Phrases that are preceded by neither are thus not inflected for gender and cannot be 'wrong'. It is therefore sensible to exclude these 'unmarkable' elements from an error analysis, since what we are interested in is the proportion of errors our participants make on phrases that require marking.

We have applied a fairly fine-grained classification of different types of pronouns and determiners. 'Pronoun' is an overarching category for eight different types of pronouns listed here in alphabetical order:

demonstrative pronouns: *der, diese, dieses* etc. (this, that)
indefinite pronouns: *keine, keiner, niemand* etc. (nobody, none)
interrogative pronouns: *wer, wem, was* etc. (who, whom, what)
personal pronouns: *ich, sie, wir* etc. (I, she, we)
possessive pronouns: *mein, ihr, unser* etc. (my, her/their, our)
reciprocal pronouns: *sich, einander* (each other)
reflexive pronouns: *mich, sich, uns* etc. (myself, herself, ourselves)
relative pronouns: *der, dem, deren* etc. (who, whom, that)

This list illustrates that it is best, when establishing categories for coding, to include all possible categories, irrespective of whether they will eventually be used in the analysis, since it is much easier to conflate categories after the fact than to go back and recode further distinctions.

Apart from coding the type of pronoun, we also manually included information about the noun referent of the pronoun. This is important in order to determine whether pronouns are being used in a target-like manner by native and non-native speakers alike. Even highly proficient speakers who have knowledge of the grammatical gender of a particular lexical item might lose track of which pronoun they have to use when it occurs at a long distance from its antecedent (Hammer et al. 2007). Consider a simple sentence like 'Sie ist heruntergefallen' (She fell down). This is a grammatical sentence in German, as long as the feminine anaphoric personal pronoun *sie* does not refer to a masculine antecedent, such as *der Ball* 'the ball'. In most cases, gender marking on pronouns is thus unambiguous. However, the sequence from the Charlie Chaplin film described above features a girl or young woman as a main character. 'Girl' is *Mädchen* in German, one of the few words in which grammatical gender and biological sex are at odds with one another, as the lexical item is of the neuter gender. This leads to pronominal references to this noun sometimes being guided by grammatical gender and sometimes by biological gender. Knowing the noun that a pronoun refers to will enable us to analyze the reference pattern that our participants prefer for this and other interesting cases.

The syntactic structure of the gender-marked elements was coded for full and bare nouns. There are five different possible structures in our coding system: (1) 'N' for bare nouns; (2) 'Det+N' for nouns preceded by a determiner; (3) 'Adj+N' for nouns preceded by an adjective; (4) 'Det+Adj+N' for nouns preceded by both

(multiple adjectives are counted as one); (5) ‘Det+Adj’ for noun phrases with an empty noun slot, which is grammatical in German.

Gender (four levels: masculine, feminine, neuter, NA = not available), number (two levels: singular, plural), case (four levels: nominative, genitive, accusative, dative) and animacy (only in pronouns, two levels: animate, inanimate) are self-explanatory labels. The ‘NA’ label for gender was required for three cases: (1) bare nouns in which gender marking is not possible; (2) plural nouns, as German distinguishes the three genders only in the singular; (3) code switches because words that are not part of the German lexicon typically do not have a conventionalized grammatical gender.

We wanted our coding system to also differentiate all types of determiners that can occur in German. This was put into practice by adding the category ‘definiteness’, which also helped to distinguish nouns that are preceded by a definite determiner from nouns that are accompanied by an indefinite article. Regardless of the type of determiner, the syntactic structure was always coded as ‘Det(+Adj)+N’ to avoid duplicating information. The six types of determiners that occur in German are (in alphabetical order):

definite: *die, der, das* etc. (the)
demonstrative: *dieser, diese, dieses* etc. (this, that)
indefinite: *ein, eine, eines* etc. (a, an)
possessive: *mein, ihr, unser* etc. (my, her/their, our)
quantifying: *alle, kein, keines* etc. (all, no)
interrogative: *welche, welches, wessen* etc. (which, whose)

Error coding was a crucial part of our coding system. We were interested in cases where speakers fall short of target-like gender marking. A complicating factor in this type of analysis is the fact that grammatical gender in German is tightly interwoven with the case and number system (see also Sect. 1.3.2). In some instances, speakers make errors that cannot conclusively be coded as gender, case or number errors. Take the following three examples:

20. *XYZ: das Milch schmeckt lecker.
21. *XYZ: dem Milch schmeckt lecker.
22. *XYZ: der Milch schmeckt lecker.
    ‘The milk is tasty.’

In all sentences, the correct form for the subject noun phrase would be ‘die Milch’, indicating a nominative (or accusative) feminine form. In (20), it has been replaced by a determiner which is ambiguous between a neuter singular nominative and neuter singular accusative. We can clearly label this as a gender error because the determiner ‘das’ can be nominative, just like the target form, and does not appear anywhere in the feminine part of the gender paradigm. Therefore the error can only come about by wrong gender assignment to the noun in question. In (21), the determiner indicates a dative form from either the masculine or the neuter class. The speaker producing this utterance gets both the gender and the case wrong. This type

of combined mistake would be labeled as both a gender error and a case error in our coding system. In (22), the determiner is gender- and case-ambiguous: the speaker might have selected the wrong case form, producing a dative or genitive instead of a nominative form in the right gender, or might have erred on the side of gender, replacing the feminine form by a masculine form in the nominative. We cannot classify this type of mistake as a case or gender error. Number errors which cannot be clearly identified were also classified as ambiguous errors.

A full line of gender coding according to our system would look like this in the incorrect sentence (21) from above, repeated here for convenience:

23. *XYZ: dem Milch schmeckt lecker.
    %fnp: ″dem Milch″ $type: FullNoun $cont: Det+N $gender: fem
    $number: sg $case: NOM $def: def $GenderError $CaseError.

We have coded the noun phrase as a feminine form in the nominative although the speaker produced a masculine or neuter form in the dative. This is because we are primarily interested in the target forms that speakers find most difficult to get right instead of analyzing which wrong forms are most frequently used instead of the correct ones. It would have been possible to code both, but as we have seen in some cases, it is impossible to classify an incorrect form, whereas correct forms are usually apparent from the context.

To conclude, some remarks on the technical implementation of a gender coding system are in order. To avoid having to type out all the coding tiers like the one in (23), macros can be used. Most text processing software allows you to record sequences of actions and play them back to repeat these actions automatically. The macro consists of several lines of code that define exactly which actions are to be performed and under which circumstances. For instance, a simple example is recording (or writing) a macro that replaces every comma in a text by a semicolon under the condition that the comma is followed by a space. Macros can also, for instance, insert a coding tier containing predefined information. Macros are typically accessed through the graphical user interface of whatever software in employed. In our case, we created buttons that triggered the playback of a macro that inserts an entire coding tier for, say, a full masculine noun in the nominative gender. Other macros, accessible in the same way, take care of editing this line—for instance, if the noun is preceded by an adjective, in a different case or incorrectly marked. Given the complexity of the gender coding tier, the use of macros both allows a significant gain in working speed and helps avoid formal mistakes such as typing errors.

Any type of text manipulation (such as adding, deleting, replacing) is possible using macros although the specifics depend on the software and programming language used. We recommend using a dedicated text editing software, such as Notepad++, which is available for free. More importantly it does not share the propensity of word processors such as Microsoft Word or OpenOffice Writer to automatically improve on text layout and structuring, such as replacing ″typewriter quotation marks″ by “typographic quotation marks” or intelligently inserting spaces

when copying and pasting text. These automatisms are likely to interfere when trying to code spontaneous speech. This means that you either have to deactivate them (making them unavailable for other work done in the same software) or, much easier, switch to software that does not do anything but edit text. Some software for transcription and coding will not interface with an editor like Notepad++. In this case, stand-alone macro software, which will record and play back actions in any application, might best serve your needs.

Suggestions for Further Reading

Kormos, J., and M. Dénes. 2004. Exploring measures and perceptions of fluency in the speech of second language learners. *System* 32(2): 145–164. doi:10.1016/j.system.2004.01.001.

Schmid, M.S., and H. Hopp. 2014. Comparing foreign accent in L1 attrition and L2 acquisition: Range and rater effects. *Language Testing* 31(3): 367–388. doi:10.1177/0265532214526175.

Wray, A. 1999. Formulaic language in learners and native speakers. *Language Teaching* 32(4): 213–231. doi:10.1017/S0261444800014154.

Wray, A. 2002. *Formulaic language and the lexicon*. Cambridge: University Press.

References

Ellis, N.C. 2003. Constructions, chunking, and connectionism: The emergence of second language structure. In *The handbook of second language acquisition*, ed. C.J. Doughty, and M.H. Long, 63–103. Malden: Blackwell.

Gullberg, M., and S.G. McCafferty. 2008. Introduction to gesture and SLA: Toward an integrated approach. *Studies in Second Language Acquisition* 30(2): 133–146. doi:10.1017/S0272263108080285.

Hammer, A., R. Goebel, J. Schwarzbach, T.F. Münte, and B.M. Jansma. 2007. When sex meets syntactic gender on a neural basis during pronoun processing. *Brain Research* 1146: 185–198. doi:10.1016/j.brainres.2006.06.110.

Hieke, A.E. 1981. A content-processing view of hesitation phenomena. *Language and Speech* 24: 147–160. doi:10.1177/002383098102400203.

Klein, W., and C. Perdue. 1989. The learner's problem of arranging words. In *The crosslinguistic study of sentence processing*, ed. B. MacWhinney, and E.A. Bates, 292–327. Cambridge: University Press.

MacWhinney, B. 2014a. The CHILDES project: Tools for analyzing talk—Electronic edition Part 1: The CHAT transcription format. http://childes.psy.cmu.edu/manuals/CHAT.pdf.

MacWhinney, B. 2014b. The CHILDES project: Tools for analyzing talk—Electronic edition Part 2: The CLAN programs. http://childes.psy.cmu.edu/manuals/clan.pdf.

McCarthy, P.M., and S. Jarvis. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods* 42(2): 381–392.

Noyau, C., and U. Paprocka. 2000. La représentation de structures événementielles par les apprenants: granularité et condensation. *Roczniki Humanistyczne* 48(5): 87–121.

Sanz Espinar, G. 1999. *Le lexique des procès dans le récit en espagnol et en français langues maternelles et langues étrangères* (PhD dissertation). Université Paris X Nanterre and Universidad Autónoma de Madrid.

Shriberg, E.E. 2002. To "errrr" is human: Ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association* 31(1): 153–169. doi:10.1017/S0025100301001128.

Chapter 5
Eye-Tracking and the Visual World Paradigm

Sanne M. Berends, Susanne M. Brouwer and Simone A. Sprenger

Abstract This chapter will focus on the use of eye-tracking in the visual world paradigm. This method can be employed to investigate a number of language comprehension issues, and we will begin with a brief overview of the history of the method and some of the applications. More centrally, we will discuss how it can be used to assess the impact of cross-linguistic interference, proficiency levels, and age of onset in L2 acquisition and L1 attrition, with an introduction to the issues that are involved in designing a study using this technique. As a case in point, we present and discuss the specific experiment employed within the multi-task, multi-language and multi-lab study on which this book is based, with special attention to the issues for analysis that arise when data from multiple systems must be combined.

Keywords Bilingualism · Grammatical gender · Eye-tracking · Second language acquisition · First language attrition

5.1 Eye-Movements and Cognition

The world is filled with visual stimuli which are constantly competing for our limited attentional resources. During visual exploration, we need to select and attend to those things that contain relevant information and ignore others. What we look at reveals a great deal about what is going on in our minds. Eye-tracking technology exploits this close temporal link between gaze and cognition to study the fast and highly automatic processes involved in language processing.

In 1974, Cooper was the first to track the eye movements of participants as they listened to short narratives while looking at a display of objects. He discovered that participants' eye gaze was drawn to objects mentioned in the narratives.

Electronic supplementary material The online version of this article (doi:10.1007/978-3-319-11529-0_5) contains supplementary material, which is available to authorized users.

M.S. Schmid et al., *Designing Research on Bilingual Development*,
SpringerBriefs in Linguistics, DOI 10.1007/978-3-319-11529-0_5

Participants were, for example, more likely to look at a picture of a lion when hearing the phrase '…when suddenly I noticed a hungry lion' than at a picture of a camera. Fixations were often initiated before the spoken word was even completed, indicating that visual and language processing are closely time-locked.

5.1.1 Gaze and Language Processing

The last decade has brought the technology to measure eye movements within reach for many research labs. Contemporary eye-trackers provide us with high-resolution quantitative evidence of a listener's visual and attentional processes (Duchowski 2002). However, while the field of psycholinguistics has seen a surge in the application of eye-tracking techniques to the study of reading since the 1980s (see, e.g., Clifton et al. 2007 for an overview) it took more than two decades until Tanenhaus et al. (1995) introduced its use in the field of auditory language processing. Since then, eye-trackers have frequently been used to study interactions between vision, attention, and the processing of spoken language.

The experimental paradigm introduced by Cooper (1974) and Tanenhaus and colleagues (1995) is known as the *visual world paradigm* (VWP, for an extensive review, see Huettig et al. 2011). Participants listen to a spoken utterance and simultaneously look at a visual scene containing various objects while their eye movements are monitored. The spoken utterance is usually related to one or more objects in the scene and the question is whether, and when, people look at these objects. When the time to launch a saccade (an eye-movement to another location) is taken into account, the point in time at which the listener's gaze is directed towards an object that has been named provides an excellent estimate of the time at which the word has been recognized (Allopenna et al. 1998). Manipulating the relationship between the objects and the linguistic input (e.g., making them harder to distinguish due to similar speech sounds) allows researchers to test theories about the way in which listeners access information in their mental lexicons.

This information can be deduced from the listeners' gaze patterns across time. In the VWP, the type of display can range from semi-realistic scenes (see Fig. 5.1 for an example) to arrays of objects (see Fig. 5.2).

Some objects are mentioned in the spoken utterance and are the targets, while others that overlap with the target to some degree function as competitors. Objects that are completely unrelated serve as distractors. The proportion of fixations on an object, time-locked to the auditory presentation of the target word, is taken to be an indication that (partial) lexical access has been achieved. For example, in a study by Allopenna et al. (1998), participants were instructed to listen to sentences such as 'Pick up the beaker; now put it below the diamond'. The names of some of the objects in the visual display were phonologically similar to the name of the target object. For example, the target object *beaker* was displayed with a competitor that phonologically overlapped at onset position (*beetle),* with a competitor that phonologically overlapped at rhyme position (*speaker*) and with a phonologically

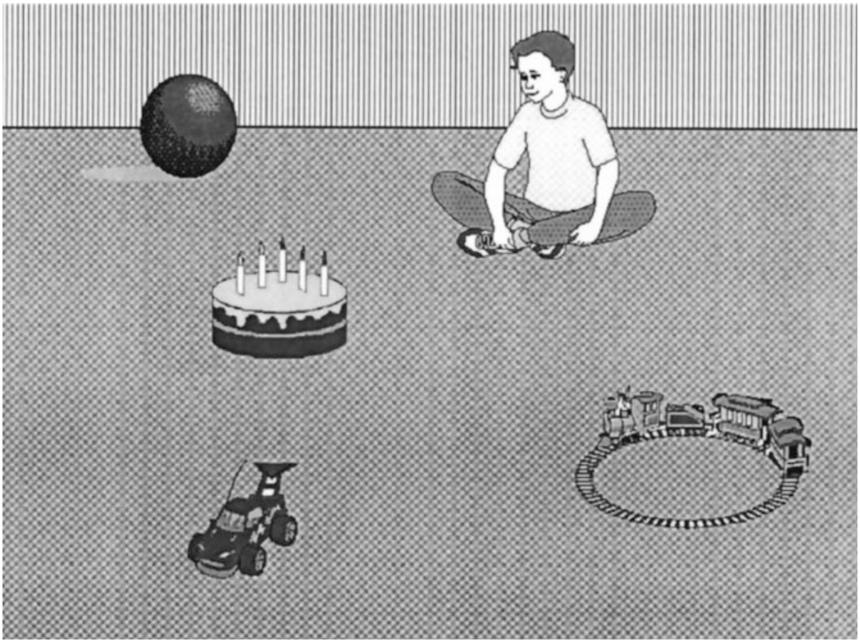

Fig. 5.1 A semi-realistic scene from an eye-tracking study (Altmann and Kamide 1999, their Fig. 1, p. 250, reprinted with kind permission from Elsevier)

unrelated distractor (*carriage)*. Results demonstrated that listeners fixated more on objects that overlapped with the target signal phonologically (*beetle, speaker*) than non-overlapping ones (*carriage*). Critically to their research question, listeners looked more often at onset competitors (*beetle*) than rhyme competitors (*speaker*), although both were fixated often enough to suggest that they were partially activated. These findings show that information at onset position is more influential in constraining lexical selection than information at final position.

The original VWP has been modulated in various ways to accommodate the demands of specific research questions. For example, McQueen and Viebahn (2007) developed a version of the paradigm in which the objects in the visual display are replaced by printed words. The benefit of this variant is that the critical stimuli do not need to be imageable, which makes it easier to design controlled sets of materials. Other changes involve the visual array, which originally contained real world objects which participants were instructed to manipulate (e.g., 'Pick up the *beaker*[1]; now put it below the diamond'), but more recently has involved objects or scenes presented on the screen with the simple instruction to look at the screen while listening to a description (e.g., 'The boy will eat the *cake'*), in order to examine effects of sentence context.

[1]Targets will be presented in italics throughout.

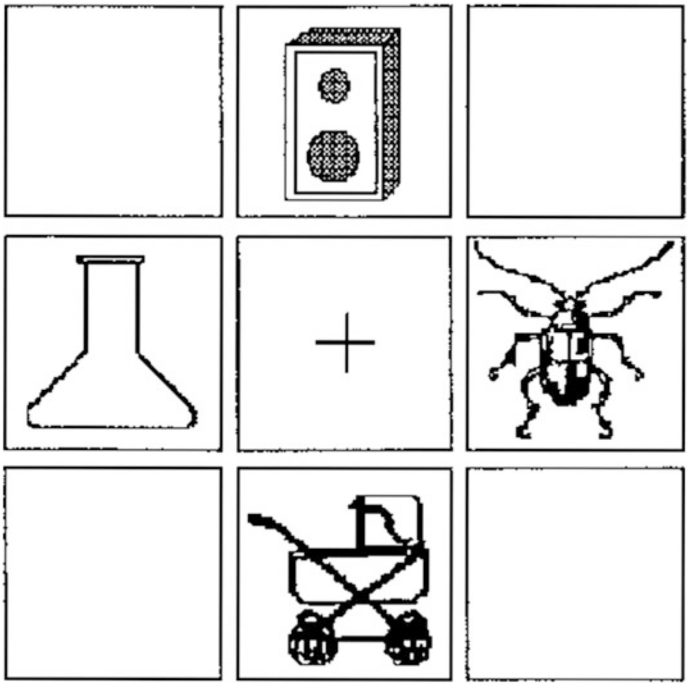

Fig. 5.2 An array of objects from an eye-tracking study (Allopenna et al. 1998, their Fig. 3, p. 427 (detail), reprinted with kind permission from Elsevier)

Until recently, the VWP was mainly used in order to answer research questions concerning monolingual language comprehension (for extensive overviews of the investigated topics in monolingual research, see Altmann 2011; Huettig et al. 2011; Tanenhaus and Trueswell 2006). Comparatively few VWP studies have been dedicated to ***L2 acquisition*** research. One of the crucial questions in this area of research is how lexical access occurs in bilinguals' native and second languages and to what extent the two languages interact during online processing. To the best of our knowledge, Spivey and Marian (1999) were the first to apply the VWP to L2 processing. In their study, Russian-English bilinguals were instructed to listen to Russian sentences (e.g., 'Poloji *marku* nije krestika'—'Put the *stamp* below the cross'). On the screen a stamp, a marker, and two unrelated distractor images were displayed. In the participants' L2, English, the object label *marker* is phonologically similar to the L1 target word *marka*, stamp. An analysis of the participants' eye movements indicated a higher proportion of eye-movements towards the between-language competitor *marker* than towards the unrelated distractors, indicating that, even if they are listening to their L1, bilinguals simultaneously activate lexical representations for words in both their L1 and their L2. The VWP can thus successfully be applied to research questions concerning L2 processing.

5.1.2 Advantages and Challenges of the Method

A first advantage of the VWP is that it is an *online* method, tapping into spoken language comprehension as it occurs and revealing aspects of processing that listeners are often completely unaware of. Second, in contrast to other (offline) methods, the listeners are not asked to perform meta-linguistic judgments that might lead to an over- or underestimation of their language abilities or encourage strategies making use of explicit knowledge. Third, the data are with a high temporal resolution (on the time-scale of milliseconds), providing precision in determining when responses to spoken language begin to differ. Fourth, the topics of interest can be investigated under natural, relatively realistic conditions in which listeners hear words, sentences, or stories that are pragmatically relevant. A last advantage is that this technique can be used for participants of most ages, since it does not require participants to read or carry out complex tasks; it has been used very successfully with preschool children in a number of studies (e.g., Lew-Williams and Fernald 2007).

A challenge lies in the fact that, as eye movements reflect the interplay of language, vision and attention, using them to study linguistic processing per se requires careful experimental design to control for possible confounds with visual processing and attention allocation. For example some colors, such as red, attract more attention than others and some positions are more frequently fixated than others, due to differences in visual processing and attention. In addition, the spoken utterances always need to be related to the visual stimuli on the screen, thereby limiting the stimulus material to either concrete, pictureable objects or relatively short written words.

5.1.3 Eye-Tracking and Grammatical Processing

The VWP has been used to investigate language comprehension on various levels of processing, ranging from discourse and sentence level to lexical processing (e.g., phonological processing and bilingual word recognition) as well as their interaction. For example, Dahan et al. (2000) used the VWP to investigate the influence of the semantic and syntactic context on spoken word recognition. Specifically, they asked whether prenominal gender cues influence the speed of spoken word recognition. In their study, French native speakers were presented with displays consisting of objects with names that shared the same phonological onset but differed in grammatical gender (e.g., $bouteille_{FEM,}$ bottle; $bouton_{MASC,}$ button). When nouns were preceded by a gender-ambiguous plural article (e.g., 'cliquez sur les_{AMB} *boutons'*, 'click on the buttons') listeners were equally likely to fixate the target and the phonological competitor (e.g., *bouteilles*, bottles) directly following the presentation of the first few phonemes. However, when a gender-marked article preceded the noun (e.g., 'cliquez sur le_{MASC} *bouton'*, 'click on the button'), listeners more rapidly fixated the correct target and the gender-marked article eliminated the

interference of the phonological competitor. These findings demonstrate that gender cues earlier in the sentence can minimize the set of possible candidates for a target noun and therefore facilitate language processing.

The same effect has since been found for article-noun combinations whose onsets did not overlap (Paris et al. 2006 for German; Lew-Williams and Fernald 2007 for Spanish, Loerts et al. 2013 for Dutch and Hopp 2013 for German). Lew-Williams and Fernald (2007) showed that the gender anticipation effect can even be found in preschoolers.

A question for L2 acquisition is whether L2 learners are, in principle, able to exploit gender marking in a similar way. This is an important question in the context of the debate on ultimate attainment among L2 learners. Of particular interest is the question of which conditions lead to nativelike processing and what role AoA and the presence and/or similarity of the gender system in the L1 play for the facilitation effect in L2. Thus far, results have been mixed, even when the same L2 is examined. Lew-Williams and Fernald (2010) and Grüter et al. (2012) did not find gender anticipation or facilitation effects for intermediate and advanced English learners of Spanish, despite the fact that Spanish has quite a transparent gender system, suggesting that a non-gender L1 precludes the acquisition of native-like L2 gender processing routines. In contrast, Dussias et al. (2013b) did find gender facilitation effects in highly proficient English speakers of Spanish. In fact, their results suggest that the presence of a gender system in the L1 may interfere with gender facilitation in the L2: A population of Italian L2 learners of Spanish that was also tested only exploited the gender cue on feminine articles and not on masculine articles. This may be explained by the fact that a greater percentage of the masculine items had opaque gender, whereas in general gender in Spanish is transparent. Another factor might have been the difference between the definite article systems of the two languages. Whereas Italian has two masculine definite articles (*il* and *lo*) and one feminine definite article (*la*), Spanish only has two definite articles (*el* for masculine and *la* for feminine).

Hopp (2013) investigated native English learners of German, which has a non-transparent gender system (see Chap. 1), and found anticipatory effects, suggesting that native-like performance on gender can be found without a gender system being present in the L1 and in the absence of phonological or form regularities of gender agreement, like those characteristic of Spanish. Loerts et al. (2013) investigated gender processing in Slavic learners of Dutch, a language which is also non-transparent. Their findings, like those of Dussias et al. (2013b), suggest that a different expression of gender in the L1 might modulate the degree to which gender marking can be used in L2 processing, since Polish learners of Dutch, who encode gender in the L1 but not on articles, showed no effect of gender facilitation.

To conclude, the factors that govern the extent to which L2 learners can acquire gender anticipation and facilitation effects during spoken language comprehension are not yet clear. While the studies that have been conducted so far suggest that it is indeed possible for L2 learners to make use of gender agreement for these purposes under certain circumstances, it is as yet unclear in which way the interaction of a range of predictors can affect the outcome. These predictors include characteristics

of both the first and the second language, the level of proficiency in the second language, the AoA of the L2 learners, and potentially others.

In this context, it is also interesting whether such anticipation and facilitation effects are stable in a native language under conditions of language attrition. To our knowledge, no previous studies exist which address gender anticipation and facilitation in L1 attriters. The findings reviewed above suggest that there may be complex interactions between the bilingual's languages that determine whether cues are used to facilitate the access of upcoming information. In order to gain further insight into how gender processing is affected by bilingual development, and to elucidate the impact of the predictors named above, it is important not to confine the investigation to the later-learned L2. By comparing L2 learners with L1 attriters (individuals who have become highly proficient or dominant in the L2 after an extended period of immersion), the impact of some of these predictors—for example, proficiency and dominance—can be disentangled from others—for example AoA and the order of acquisition (see Chap. 1 for a more complete discussion). We propose that deeper insight into multilingual grammatical processing can only be gained on the basis of comparisons among participants that vary along all of these dimensions. Such investigations should thus include native controls, L2 speakers and L1 attriters from different linguistic backgrounds across a range of proficiencies and AoAs. In the following sections we will discuss the considerations that went into constructing a study on gender processing designed to achieve this goal. However, most of the considerations are sufficiently general to apply to other studies using this paradigm.

5.2 General Design Issues

In this section we present a number of considerations which should be carefully taken into account when designing a VWP experiment and discuss some of the strategies that have been used to deal with each of them.

5.2.1 Fixating Visual Objects: Important Potential Confounding Factors

The types of visual displays that have been used in VWP experiments vary depending on the research question. In general, two different types of displays can be distinguished. The first type consists of arrays of line drawings (black and white or colored), or of pictures of real objects. The second display type is made up of semi-realistic scenes, which consist either of drawings of pictures presented on a computer screen or of real objects laid out on a workspace (see Figs. 5.1 and 5.2). The main difference between the two paradigms is that semi-realistic scenes give a more natural context in which the impact of world knowledge can be tested

(Henderson and Ferreira 2004), whereas in the paradigm with arrays of objects the influence of world contextual knowledge is reduced to a minimum, providing the opportunity to isolate the activation of conceptual and lexical information associated with individual words stored in the mental lexicon (Huettig et al. 2011).

Picture selection is a crucial step in developing a VWP experiment. The objects to be used in a visual display can be selected from picture databases such as the one created by Snodgrass and Vanderwart (1980). This database comprises 260 black and white line drawings which have been normed for name agreement, image agreement, familiarity and visual complexity. Since these factors might potentially influence the eye gaze, they should be taken into account when selecting the visual stimuli, making pre-normed stimuli an excellent choice (note, however, that factors such as familiarity may vary depending on the culture—an important consideration for investigations of multilingual development!).

Pictures of real objects are more detailed than black and white drawings, and can therefore enhance recognition and facilitate naming agreement. However, visual complexity differences between pictures of real objects are more difficult to control: Some pictures will be more easily recognized than others (Dussias et al. 2013a), which encourages earlier fixations when the objects are named. As this visual complexity bias might obscure any potential anticipation effect, we suggest using line drawings from the Snodgrass and Vandewart (1980) picture database in order to keep the variance in the complexity of the pictures as low as possible. As those pictures have only been normed for English naming agreement, studies investigating other languages should pilot the pictures to be used on native speakers of the languages represented in the experimental population, in order to ensure that each picture stimulus will indeed elicit the intended nouns.

The Snodgrass and Vandewart pictures are ***copyright-protected***, and in order to use them in any study or for publications it is necessary to obtain a license. Other databases of images are available (for a list of suggestions see http://www.cogsci.nl/stimulus-sets). Whatever images are selected for any given study, it is of vital importance that the researcher should consider and explore any potential copyright issues, since infringement of such rights can have serious consequences (and also make it difficult if not impossible to publish the results from the study).

Equally important is the consideration of how the ***spoken utterances*** that participants will hear may constrain the visual characteristics of the objects being employed. For example, in a study which investigates the use of gender in anticipation, it is important to reduce effects of sentence context that might also lead to anticipation because the target noun is semantically the most likely object to fit a given sentence frame. Many VWP paradigm studies therefore opt for minimally constraining contexts and present only noun phrases consisting solely of the determiner and the noun (e.g., *de appel*, 'the apple'). However, this practice is problematic for two reasons. First, since launching an eye-movement takes ca. 200 ms (Matin et al. 1993), and since determiners are considerably shorter than that in many languages, such phrases may not allow participants enough time to translate their anticipation of the upcoming noun into an actual saccade. Second, it has been proposed that L2 learners acquire frequent combinations such as

determiner plus noun as chunks, and that they therefore might show a preference for the target, but for reasons that are unrelated to gender as a structural property of the noun. It is therefore preferable to extend the noun phrase by an intervening adjective (‘the ADJ apple’), where the structure of the language allows this.

This raises another problematic question, namely what type of adjective should be used: Evaluative words, such as ‘pretty’, ‘nice’ etc. may confuse participants who do not agree with the assessment—or where the description may better apply to other objects. Depicting adjectives that refer to more objective properties of the target (‘large’, ‘heavy’ etc., see Paris et al. 2006) may enhance or reduce the noticeability of the target (which then also has to be larger or heavier than the other items in the array) and thus confound the anticipation effect.

An alternative solution is suggested by Loerts et al. (2013), who investigated whether Dutch native speakers use gender marking to predict the correct referent. The visual display used in this study contained a target, a competitor and two distractors, all of them represented as colored line drawings. The competitor was either the same or a different gender and/or color as the target (while the distractors were always represented in different colors and had different genders), and the intervening adjective named the color. For example, one such array depicted a red apple (common, target), a red cake (common, competitor), a yellow lock (neuter, distractor) and a blue book (neuter, distractor). In this case, when the participant heard the phrase ‘click on the$_{\text{COM}}$ red…’, both target and competitor were equally likely to be fixated until the onset of the noun (apple vs. cake), since both were potential referents of the noun phrase. In an array where the two red objects did not share their gender and the competitor was, for example, a red book, participants were able to differentiate them at an earlier stage, despite the fact that both were represented in the same color.

Results showed that the color of the pictures interfered with the gender anticipation effect to some extent: Participants’ fixations to targets were initiated later when the target was brown as compared to other colors (red, yellow, blue, green). This finding indicates that some colors are more salient than others and therefore attract more attention. Loerts et al.’s (2013) findings revealed that the effect of the color manipulation was stronger than the gender facilitation effect in those visual displays in which the target and competitor shared color.

While inserting color adjectives between determiner and noun is therefore probably the best way of constructing noun phrases that allow the participant enough time to use gender agreement information encoded in the determiner, the color interference may override the gender facilitation effect. This, however, can be eliminated by presenting all objects within the same VW display in the same color (while making sure that color is counterbalanced across conditions).

Another aspect of the design that needs attention concerns the number of objects and the layout of the visual displays. Both of these factors can also influence the time it takes to fixate the correct object. A first factor to consider is the number of object positions on the visual display (Ferreira et al. 2013, see Figs. 5.3 and 5.4). Ferreira and colleagues studied the influence of the complexity of the visual display on the interpretation of garden-path sentences like ‘Put the book on the *chair* in the

Fig. 5.3 The visual display used by Ferreira et al. (2013), their Fig. 1 (reprinted with kind permission from Elsevier)

Fig. 5.4 The visual display used by Ferreira et al. (2013), their Fig. 4 (reprinted with kind permission from Elsevier)

bucket.' Here, the prepositional phrase *on the chair* is temporarily ambiguous because it can either be interpreted as being the goal (the location where the book is to be placed) or the modifier (the location from which it is to be removed, in order to be placed in the bucket). In the unambiguous equivalent 'Put the book that's on the *chair* in the bucket', the same prepositional phrase *on the chair* can only be interpreted as the modifier. In a visual context with a book on a chair, a single book, an empty chair and an empty bucket participants were less likely to fixate the empty chair even in an unambiguous sentence. This can be interpreted as suggesting that when two books are present in the display, the listener is more likely to assume that the chair serves as a location, identifying the appropriate book, i.e. that the modifier interpretation of the prepositional phrase is favored.

However, when the single book is replaced by an unrelated object, garden path sentences typically elicit more gazes towards the incorrect goal, that is, the empty chair (Tanenhaus et al. 1995) as compared to the unambiguous counterpart. Ferreira and colleagues also used instructions which did or did not contain garden-paths. In addition, the complexity of the visual displays was manipulated by presenting participants with 4 (Fig. 5.3) or 12 (Fig. 5.4) objects. The findings indicated that it is more difficult to construct an online interpretation when the visual display is more complex (i.e. contains more objects). Looks to the target (i.e., the book on the chair) increased much later, which suggests that the visual search takes longer in complex visual contexts. Furthermore, the classical garden-path effect disappeared due to the delay. This suggests that it is undesirable to use an unnecessarily complex display. In order to ensure comparable results between eye tracking studies with similar research questions, it is therefore advisable to use the same number of objects in the visual display as other studies in the field.

Participants also have clear preferences for particular screen positions, depending on the reading direction in their native language. Readers whose script runs from left-to-right and from top-to-bottom will have the tendency to first direct their gaze to the upper left corner of a visual display. To control for this bias, the positions of the objects on the visual display should always be counterbalanced so that each condition (target, competitor, distractors) is presented equally often in each position and that each individual object is displayed in all positions and all conditions. To avoid repetitions of the same array for the same participant, the various uses of each object can be distributed across different lists (so that, for example, in one list, *apple* is used as target, in another as competitor, and so on); at the level of the entire experiment, the effects of position should then be balanced.

5.2.2 Presenting Auditory Stimuli: Important Potential Confounding Factors

In addition to the effects of the visual display, eye movements can also be influenced by inadvertent properties of the ***auditory stimuli***. When selecting the critical

words for an experiment on gender processing, several important factors must be controlled in order to avoid confounds in the experiment. One of these factors is ***phonological overlap*** between the target object and the other objects presented on the same visual display. Recall the experiment by Allopenna et al. (1998) described earlier, in which participants were found to direct their eye gaze more towards the phonological onset competitor than to the rhyme and unrelated competitor. However, participants also fixated the rhyme competitor more than the unrelated competitor. Studies investigating anticipatory gazes should therefore avoid phonological overlap at either onset or rhyme as much as possible.

Furthermore, ***word frequency*** can also modulate lexical access speed and therefore impact on fixation time. For example, presenting a comparatively low-frequency auditory target like *bench* alongside a high-frequency phonological competitor (e.g., *bed*), a low-frequency phonological competitor (e.g., *bell*) and an unrelated distractor elicits more fixations to the high- than to the low-frequency competitor (Dahan et al. 2001). Importantly, this effect is not limited to cases with phonological competition. Dahan and colleagues also found frequency effects in eye movements toward targets without phonologically related competitors. Eye gaze latencies towards targets with high frequency names (e.g., *horse*) were faster than for targets with low frequency names (e.g., *horn*). It is therefore extremely important that the lexical frequency of all items in an array be stringently controlled. For studies of L2 acquisition, it may furthermore be advisable to select only items of comparatively high frequency, since this will reduce the chance of participants being unfamiliar with certain words or their gender.[2]

When carrying out an L2 acquisition study, cross-linguistic ***gender overlap*** of the target noun can also potentially affect the results. Weber and Paris (2004) presented French learners of L2 German with German instructions to click on a target object, and the target noun was preceded by a gender-marked article. Participants saw visual displays with a target and a competitor, as well as two distractors. The target and the competitor always had phonologically overlapping onsets not only in the language of the experiment (German, e.g. *Perle* 'pearl' vs. *Perücke* 'wig') but also in their French translation equivalents (*perle, perruque*). In addition, whereas the gender of the target was always shared between the German noun and the French equivalent (e.g., die_{FEM} *Perle*, la_{FEM} *perle*), the competitor's gender was manipulated in such a way that it either had the same (e.g., die_{FEM} *Perücke*, la_{FEM} *perruque*) or a different gender (e.g., die_{FEM} *Kanone*, le_{MASC} *canon*) across both languages. In the latter condition, where the competitor had the same gender as the target in the language used (German), earlier fixations on the target could only be ascribed to the influence of the participants' first language, French. Such a control condition is thus necessary to ensure that anticipatory effects are, indeed, based on L2 gender only. The influence of crosslinguistic competition was confirmed by the finding that Weber and Paris' French learners of L2 German only

[2]Irrespective of the frequency of the chosen items, it is highly advisable to ensure that all participants know all of the words, see Sect. 5.3.3.

showed an effect of competition when the gender of the competitor matched that of the target in *both* languages. The result suggests that participants were unable to eliminate the L1 gender while listening to instructions in the L2. Neglecting to include this manipulation in the design would therefore have produced misleading results.

5.2.3 Controlling Timing

One of the advantages of the VWP is that it allows tracking eye movements over time, to see how and when the auditory stimuli direct attention toward elements of the visual display. The fine-grained temporal resolution of eye-movements makes ***accurate timing of the recordings*** of these stimuli essential across multiple trials. This is a more challenging task for experiments using spoken language than for designs that rely on written language (see also Chap. 6). We recommend a procedure in which all stimuli (which for the purpose of the gender anticipation task take the form of sentences such as 'click on DET ADJ NOUN' or 'where is DET ADJ NOUN') have the same timing across relevant regions. To achieve this goal, we used the following procedure: Each sentence was recorded three to five times by a female native speaker who spoke a standard version of the target language and had considerable elocution training. From these recordings, the best exemplar for each stimulus sentence was selected, based on normal speaking rate and naturalistic prosody (stress, rhythm and intonation).

For the purposes of the gender experiment, there are three regions preceding the noun in the stimulus sentence: verb plus preposition (preamble), determiner (moment at which relevant information is presented), and adjective (intervening region). We established the average duration of all of these regions and subsequently adjusted them to this average length in every one of the recordings, so that the onset of the noun always occurred at exactly the same moment in each sentence. The purpose of this adjustment was to avoid a larger facilitation effect in those sentences that were pronounced at a slower rate and thus reduce between-item jitter in timing as much as possible.

A second important issue to keep in mind in relation to timing is when the auditory and the visual stimuli should be presented relative to each other. In the VWP, the presentation of the visual display often starts at or shortly before the onset of the utterance. Previous work has shown that the amount of ***preview time*** can affect the likelihood of fixations to particular objects (Huettig and McQueen 2007; Ferreira et al. 2013). Ferreira and colleagues manipulated the amount of preview time in their study on the interpretation of garden path sentences discussed above. They found the classical garden-path effect when the visual display contained four objects and was presented three seconds before the instructions initiated. However, the effect disappeared when the visual and auditory information were

presented simultaneously (i.e., preview time was reduced to zero). The authors suggest that preview time may allow participants to build better expectations of what may be referred to in the upcoming utterance. The primary goal of the preview time should be to allow the participants to construct a spatial representation of the VW array, as longer times may lead to strategies and expectations that influence the results. The ability to predict upcoming information might also increase over the course of an experiment as experience in the task grows. We therefore recommend as short a preview time as is consistent with the participants being able to carry out the task. In our case we chose to give no preview time, the most extreme option, as both our visual display and our instructions were comparatively simple.

5.2.4 Summary of General Considerations

In sum, we recommend a research design that takes into account the following factors:

- properties of the lexical items: word frequency (based on large speech/written corpora), phonological overlap (at onset or at rhyme) between targets, competitors and unrelated items, controlled gender overlap between items in the target language and in the L1 of L2 speakers where applicable (in our case Polish and Russian)
- properties of the picture stimuli: reliability of naming, effects of color, size and complexity
- properties of the auditory stimuli: equal duration of the relevant regions preceding the noun (e.g. imperative/preposition, determiner, adjective) for all recorded stimuli, achieved through adjusting the length of the recorded segments
- properties of the visual world scene and presentation: locations of the target and competitor objects counterbalanced across the regions of the scene which are employed, a preview time that is consistent with the participants being able to carry out the task.

These design principles are quite general and apply to both the Dutch and the German version of the experiment we present below. However, the actual implementation of the design differed slightly with respect to the target nouns, carrier sentences and counterbalancing results due to differences in gender and case systems between Dutch and German, a problem that will occur in any multi-language study. In the following, we therefore provide descriptions of how the general design was adjusted for the Dutch and German stimuli, while maintaining enough similarity for comparison across languages.

5.3 The Present Experiment

5.3.1 Rationale of the Experiment

In our study, we were interested in the effects of the characteristics of both a bilinguals' languages on gender processing. In addition we were interested in how the salience of the gender system affects L2 acquisition and L1 attrition: While both Dutch and German have opaque gender systems, which are generally not predictable on the basis of morphological or phonological characteristics of the word, gender is more salient in German due to the interaction with case marking (see Chap. 1 for a full discussion). We therefore compared adult learners of Dutch and German with various types of L1, L1 attriters and predominantly monolingual speakers of these languages. Our goal was to assess the effects of language dominance, order of acquisition, proficiency and AoA on the use of gender agreement to facilitate information retrieval during online processing. In order to measure the listeners' response to grammatical gender information in Dutch and German noun phrases, we manipulated gender overlap between the objects displayed, as in the studies cited above.

In each trial participants saw four line-drawings on a screen and heard a sentence directing their attention to one of these objects (e.g., 'Klik op het$_{\text{NEUT}}$ groene blad$_{\text{NEUT}}$', 'Click on the green leaf'). Crucially, in this sentence the gender-marked determiner must agree with the gender of the noun; if listeners are able to anticipate the correct noun based on the gender information provided by the determiner, this demonstrates a quick and automatic use of gender information in comprehension. Two types of display were used, one where there was no gender competitor (see Fig. 5.5, Panel A for an example from the Dutch experiment) and one that contained two potential target objects, the actual target and a gender competitor (see Fig. 5.5, Panel B). The difference between the two types of displays represents the within-subject factor ***gender competition***.

If listeners make use of the gender information encoded in the definite determiner to direct their gaze to the target as quickly as possible, they should settle on the *leaf* sooner in panel A than in panel B, since there is no competition from a gender-congruent competitor (in panel B, *eye* represents such a competitor). Such a difference in the visual selection of the target object can be taken as evidence for active use of gender information in comprehension. In order to get reliable estimates of the time it takes to select the visual targets, the effects have to be independent from the specific materials used. Therefore, the visual and the auditory stimuli, as well as the visual displays, had to be selected and created with care.

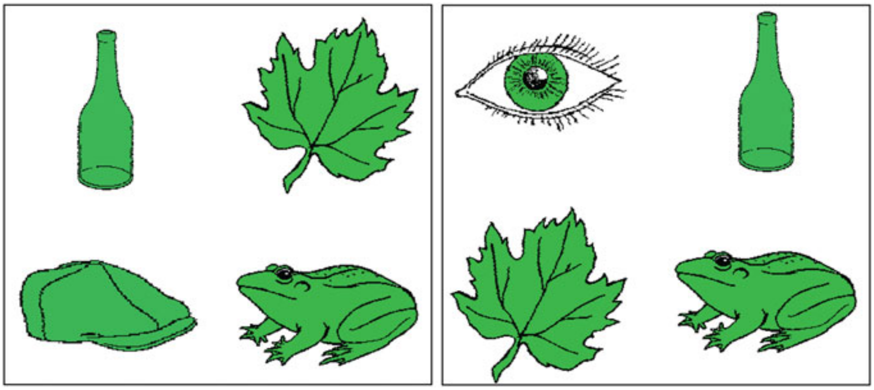

Fig. 5.5 Example of two visual stimulus arrays in the gender competition task for the spoken Dutch stimulus 'Klik op het$_{\text{neu}}$ groene *blad*$_{\text{neu}}$' (Click on the green leaf). In panel A (*left*) the objects other than the target provide no gender competition as they are all common gender, whereas in panel B (*right*) one of the objects (i.e., het$_{\text{neu}}$ oog$_{\text{neu}}$, 'the eye') is a gender competitor for the target

5.3.2 Materials

48 highly frequent Dutch nouns and 48 highly frequent German nouns referring to pictureable objects were selected as targets. While the Dutch nouns were evenly spread over the two genders encoded in the language (e.g., common: *de emmer* 'bucket', 24 items; neuter: *het potlood* 'pencil', 24 items), the German nouns were limited to two of the three genders (e.g., masculine: *der Hut* 'hat', 24 items and neuter: *das Kleid* 'dress', 24 items). No feminine German nouns were included, as the feminine singular definite article *die* is ambiguous with the plural definite article for all genders. The Dutch nouns were selected based on the Spoken Dutch Corpus' estimates of frequency of occurrence (Oostdijk 2000) and the German nouns were selected from a basic vocabulary list for German learners (Oehler and Gerretsen 1973). This was the first within-subject factor in the experiment: ***grammatical gender***.

Two noun phrase constructions were created: (1) definite determiner-adjective-noun (e.g., *het gele hek* 'the yellow fence'; *der blaue Hut* 'the blue hat'), and (2) indefinite determiner-adjective-noun combinations (e.g., *een geel hek* 'a yellow fence'; *ein gelbes Kleid* 'a yellow dress'). This was the within-subject factor ***gender structure***. In the definite noun phrase condition, gender information is marked on the definite determiner in both languages (e.g., *de*$_{com}$ vs. *het*$_{neu}$; *der*$_{mas}$ vs. *das*$_{neu}$). In the indefinite noun phrase condition gender is carried by the adjective (e.g., *gele*$_{com}$ vs. *geel*$_{\text{neu}}$; *blauer*$_{masc}$ vs. *blaues*$_{neu}$). The adjectives used were *yellow*, *blue*, *green*, and *brown*. The Dutch noun phrase combinations always followed the sentence frame 'Klik op…' (Click on…). However, it seemed best to present the German nouns in the nominative case form, since this is considered to be the default form and is the more frequent one, to make the experimental sentences more

comparable cross-linguistically. Therefore, the German noun phrases were embedded in the Wh-question 'Wo ist…' (Where is…), since the German equivalent of 'Click on' would require the accusative form. The distribution of gender across the objects on the experimental trials was either 1:3 (i.e., the gender of the target versus the gender of the unrelated competitor and both distractors) or 2:2 (i.e., the gender of the target and the related competitor versus the gender of the distractors).

The Dutch sentences were recorded by a female native speaker of Dutch living in the Netherlands and the German sentences were recorded by a female native speaker of German living in Germany. Both speakers had professional experience as radio presenter or voice over. Recordings (16 bits, 44.1 kHz) were made in a sound-attenuating double-walled booth. The duration of the preambles, the determiner and the adjective were measured and averaged across all sentences in PRAAT (Boersma and Weenink 2012). All auditory stimuli were then manipulated in Adobe Audition© 3.0 to adjust the duration of the three regions from the onset of the gender cue to the following values: determiner (Dutch: 858 ms; German: 445 ms), adjective (Dutch: 974 ms; German: 660 ms) and noun (Dutch: 1351 ms; German: 1060 ms). The stimuli were also equalized to the same rms level of 65 dB and modified to fade in and out by means of PRAAT.

In addition, for each language 12 pictures of naturally red-colored objects were selected and used to construct 24 filler displays. In contrast to the experimental displays, the filler displays contained a target, two objects with the same grammatical gender and only one of a different gender. They were included to prevent participants from strategically looking for the odd one out, since the experimental target item was always different in gender from at least two and sometimes three of the other objects. In the filler displays the target had two same gender competitors and the distractor was the odd one out. Finally, six Dutch and five German practice displays were constructed in different colors.

The words used in the two language experiments were paired with 48 pictures selected from the Snodgrass and Vanderwart (1980) standardized picture set of black and white line drawings. Color was then added to the line drawings. The possible word-picture pairs were constrained, as they must plausibly appear in either yellow, blue, green or brown in the real world (in addition to the 12 naturally red fillers) and the pictures were colored accordingly (e.g., *sun* was colored yellow; *apple* was colored green). A naming pretest, in which participants were asked to name individually presented pictures on a computer screen, confirmed that the pictures were highly recognizable and elicited the correct target noun. Objects that did not consistently elicit the expected noun were revised or excluded. All pictures are included in the online supplementary material.

With these 48 pictures, 96 array combinations were created by positioning the objects in two-by-two grids (see Fig. 5.5); there were two versions of each array which varied as to competitor (gender matched or not); these were counterbalanced across lists to prevent repetition effects. The four objects included a target, a competitor and two distractors. All four objects always had the same color so that color did not add any information for identification of the target; nor would one of

the objects be more noticeable due to its color. The objects in the display were positioned equidistant from the center of the screen. The competitor's gender was either congruent or incongruent with respect to the target object (e.g., Dutch *vliegtuig*$_{\mathrm{neu}}$ 'airplane' vs. *lepel*$_{\mathrm{com}}$ 'spoon' for the target *potlood*$_{\mathrm{neu}}$ 'pencil'; German *Boot*$_{\mathrm{neu}}$ 'boat' vs. *Löffel*$_{\mathrm{masc}}$ 'spoon' for the target object *Haus*$_{\mathrm{neu}}$ 'house'). The genders of the distractors were always different from that of the target. In order to exclude the possibility of phonological competition, the onset of the target item never overlapped with the onset of any of the other object names in the Dutch experiment. However, in the German experiment, we could not avoid phonological overlap completely (16 items in list 1 and 18 in list 2). The lists of array combinations of targets, gender congruent and incongruent competitors and distractors are included in the online supplementary material.

For each language two pseudo-randomized item lists were created in which targets were counterbalanced to appear either with a gender-congruent or a gender-incongruent competitor, and each participant was assigned to one list. On each list, the positions of target objects on the screen were counterbalanced. That is, the target appeared with equal probability in the four quadrants of the screen over the course of an experimental run. There was no repetition of colors on subsequent trials. Both lists were divided into two blocks, so that per block each picture appeared once as target, once as competitor and twice as distractor. One block contained the auditory stimuli in which the determiner was definite, and one had the indefinite constructions.

To sum up, each participant saw 96 target arrays and 24 filler arrays, in a factorial design varying

- ***grammatical gender*** (two levels: common vs. neuter for Dutch and masculine vs. neuter for German),
- ***structure*** (two levels: definite vs., indefinite), and
- ***gender competition*** (two levels: competitor vs. no competitor)

combining to 9 conditions, each with 12 exemplars per participant.

5.3.3 Procedure

Before beginning the experiment proper, participants were presented with a series of two pictures on a computer screen and simultaneously heard a bare noun corresponding to one of the pictures presented. Their task was to click on the corresponding picture on the screen. This task ensured that all participants were familiar with the names of all objects. Participants heard only the bare noun in order to avoid any priming of the article.

For the eye-tracking experiment, participants were seated in front of a computer screen while their eye movements were monitored by an eye-tracker. The presentation of the auditory and the visual stimuli was controlled with E-Prime (Schneider et al. 2002), which could be used at all labs where data was collected. Prior to the

experiment a 9-point calibration procedure was performed; by directing the participant's gaze to specific regions of the screen, the eye-tracker is able to check whether the data being gathered corresponds to the actual location which the participant should be fixating to establish accuracy. Participants were then instructed to use the computer mouse to click on the object in the visual display representing the target item they heard in the sentence. They were asked to respond as fast and as accurately as possible. Each trial started with a central fixation cross, displayed for 500 ms in order to avoid baseline effects, followed by a visual display with four objects. The spoken sentence started simultaneously with the onset of the display. When participants clicked on an object, they initiated the next trial. Both reaction times and eye gaze data were recorded.

Each participant saw two stimulus blocks, with either definite determiners or indefinite determiners used in the auditory descriptions. The order of blocks was counterbalanced over participants. Prior to each block, participants performed a practice session of two or three trials to become acquainted with the (change in) task and procedure. Each block lasted approximately 10 min and participants were given a short break between the two blocks. The total duration of this testing session was approximately 30 min.

5.4 Data Recording and Analysis

5.4.1 Eye-Tracking Devices

As will usually be the case in studies aiming to collect data on different languages and from different populations, and therefore at different testing sites, the systems available for the experiment differed between locations and labs (see Chap. 3). For this experiment, data were obtained from SMI, SR Research and Tobii eye-trackers, each of which generates different formats of output, as will be discussed in more detail below. In addition, the physical set-up experienced by the participants differed from head-mounted through desktop-mounted to built-in eye-tracking systems. A full list of the variants and the populations which were tested on each set-up is given in Table 5.1.

Some technical differences notwithstanding, all of these systems are designed to answer the simple question "when does a participant look where?" That is, they provide us with x-y-coordinates of the participants' gaze, measured at a high temporal resolution (60–500 Hz, see Table 5.1), time-locked to the spoken stimuli. All systems analyze the reflection of infrared light from the eye in real time in order to determine the location of the pupil as well as the corneal reflection. The combination of these parameters allows the software to determine the participants' direction of gaze. Even though each set-up differed with respect to the position of the infrared emitter and the camera (and therefore with respect to its sensitivity for particular forms of movement artifacts), each of them has reliably been applied in

Table 5.1 Sampling rates of the various eye-trackers used in our project

Eye-tracker	Location	Sampling rate (Hz)	Set-up	Participant group
Eyelink II	Chicago	250	Head-mounted	Dutch attriters
	Toronto	250	Head-mounted	Dutch and German attriters
Eyelink 1000	Hamburg	250	Remote	German learners and natives
	Leiden	500	Remote	Dutch learners and natives
	London (ON)	500	Desktop-mount	Dutch attriters
SMI	Berlin	60	Remote	German learners and natives
Tobii T60	New York	60	Remote	German attriters
Tobii T120	Groningen	120	Remote	Dutch learners and natives

experimental settings in which timing information is essential. Given our within-subject design, and given our relatively large regions of interest, possible differences between eye-trackers were not expected to affect the results.

All eye-trackers were linked to and controlled by the same experiment software using somewhat different interfaces (E-Prime software, Psychology Software Tools, Pittsburgh, PA). For example, communication between E-prime and the Tobii eye-trackers was established by a set of software extensions that link the eye-tracker server with E-prime. The E-prime extensions for the Tobii package can be obtained from the Psychology Software Tools, Inc website (http://www.pstnet.com/downloads/eet/EETVersion6.pdf). The procedure built in a check for accuracy which helped ensure that the data from different set-ups was comparable; each recording session started with a calibration procedure that mapped the signal of the eye-tracker to the dimensions of the computer screen and the visual field of the participant.

5.4.2 Dependent and Independent Measures

For our project, blinks and saccades were discarded, as the focus was on fixations. Specifically, we were interested in the time course along which the listeners' gaze was directed at various objects in the visual scene. To operationalize this, we created four different visual regions of interest by splitting the screen into four quadrants, containing the four objects. Subsequently, we computed the average probability with which each object/region of interest was fixated within a given time bin (bin size of 50 ms), across a time span beginning 200 ms after the gender cue was heard. This starting time was chosen because of estimates that it takes approximately 200 ms to program and launch a saccadic eye movement (e.g., Matin et al. 1993).

In the analysis, objects were classified as targets, competitors and distractors. For the stimuli with gender marking on the determiner, we examined gaze proportions in the time window until noun onset to see if fixations reflected gender anticipation when no competitor was present.

5.4.3 Combining Data from Different Eye-Tracking Systems

Eye-trackers provide information about gaze location across time. Typically, this information is coded in terms of x- and y-coordinates, with x and y being the pixel dimensions of the experiment screen (where [0, 0] typically identifies the upper left corner). Time is logged based on the eye-tracker's internal clock time. The accuracy with which the gaze location is determined depends on the eye-tracker's sampling rate. In visual world experiments, sampling rates typically vary between 60 (one measurement every 16.667 ms) and 500 Hz. (one measurement every 2 ms). The sampling rates of the different eye-trackers used in our project are listed in Table 5.1. Given an average fixation duration of approximately 330 ms in scene viewing, in combination with relatively large regions of interest (i.e., screen quadrants) and time bins of 50 ms, all these systems provide ample accuracy for the study of spoken word recognition.

In order to be interpretable, the eye-tracking output must contain information in addition to gaze location and time. First, there must be a way to relate the eye-tracker's clock time to the timing of the experiment (e.g., onset of presentation of the picture and/or the spoken stimulus). Second, each sample must be coded for the experimental item, the trial, and the levels of one or more independent variables. Finally, each set of samples must specify the participant, and, if applicable, experimental list.

The output formats that have been chosen by the various manufacturers differ profoundly in the way in which this information is coded. Eye link data are in European Data Format (EDF), which is a standard binary file format for medical time series data. EDF files start with a header that contains general experiment information, as well as information about the calibration procedure. The header is followed by data records, of which there are various types. For our analyses, we relied on *messages* (i.e., a timestamp and some pre-specified string that is generated by the experiment software at certain events) and *samples* (i.e., a timestamp, followed by x- and y-coordinates of one eye and a measure of pupil dilation). The messages were used to mark important events in the time course of a trial (e.g., sound onset) and to provide information about the levels of the independent variables (per trial). We combined all three types of information by means of a custom-made script that we applied to the ASCII-converted files, which re-arranged the information for easier statistical analysis. That is, each sample was coded for time, experimental list, participant, trial, item, levels of the independent variables,

and time of sound onset.[3] The data were then stored in a data table format with one row per sample and one variable per column.

Tobii gaze data are already in data table format when exported from the Tobii eye-tracker server. Each row represents a sample taken by the eye-tracker. In addition to timestamp information in milliseconds, the eye-tracker provides information about the x- and y-positions of *both* eyes, a measure of pupil dilation, and a number of parameters related to the validity of the measurement. A detailed overview of these variables can be found in the Tobii manual (http://www.tobii.com/Global/Analysis/Downloads/User_Manuals_and_Guides/Tobii_T60_T120_EyeTracker_UserManual.pdf). The sample information was automatically complemented with the relevant trial coding (i.e., the independent variables, see the section on EDF output). Importantly, the onset of the auditory instruction and visual display was marked by means of an additional variable, making it possible to align the eye-tracking measurements with the timing of the spoken and visual stimuli. The tutorial section of the E-prime extensions for the Tobii package includes instructions and examples of how to use inline scripts in E-prime to create such integrated gaze data files.

The data that were acquired by means of an SMI eye-tracker were in SMI's IDF (iView Data File) format. Similar to the EDF files, IDF files are binary files that consist of a header, followed by data records. Like EDF files, IDF files must be converted to ASCII for further processing. The actual data in IDF files can be considered a hybrid of the EDF and Tobii formats, as the data records are arranged in the form of a data table (one sample per row, marked as SMP), interspersed with messages (e.g., a timestamp in milliseconds, followed by the marker MSG and the string *start recording*). With respect to the variables, the output is similar to that of Tobii. That is, each sample includes a timestamp, x- and y-positions of *both* eyes, a measure of pupil dilation, and a number of parameters related to the validity of the measurement. In addition, trial number is included. Independent measure coding was added post hoc, by merging the corresponding E-prime output with the sample data.

Once the original data format had been standardized for all types of equipment and output files, further data processing proceeded along the following lines. First, we verified that each data file was in data table format, coded for experiment, experimental list, experimental block, participant, participant-related independent variables (i.e., first language, presence of gender in L1, age of arrival), trial, trial timing, item, and experimentally controlled independent variables (related either to our hypotheses or to counterbalancing). A further issue concerns the coding of missing data across the different file types. This is important, as each time a participant blinks (or, more accurately, each time the system loses track of the eye), the eye-tracker will report missing data. The amount of data lost to either blinks or a

[3]Recall that our audio files were standardized with respect to the length of the fragment that precedes the target.

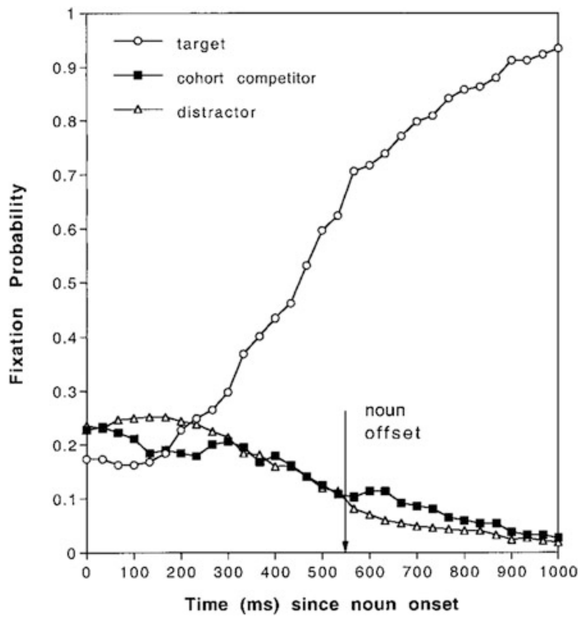

Fig. 5.6 Probabilities of fixation for the target object, the competitor object and the distractor objects (divided by two) across the time course of the noun (Dahan et al. 2000, their Fig. 5, p. 475, reprinted with kind permission from Elsevier)

poor signal is an important indicator of the overall validity of the acquired data and should therefore be subject to analysis by itself.

In order to combine data from various sources, we downsampled the gaze data to the lowest sampling rate available for each analysis. For example, if German participants were included, the lowest sampling rate was 60 Hz or one sample every 16.7 ms. Data from other eye-trackers was converted to the same rate. This step was primarily necessary for making average images of the data, since the larger bins described below for the analysis are essentially a still more stringent downsampling. The data was further minimized by excluding information that was not used in the analyses; for example, only fixations were selected and saccades and blinks were excluded. Moreover, critical items were selected and practice items and fillers were eliminated.

Third, for all trials we aligned the sound onset times to zero to establish a common time frame for all trials. Average gaze locations were aggregated per participant per trial in 50 ms time bins. Each gaze location with respect to region of interest was categorized (i.e., either one of the four quadrants of the screen or the center of the screen[4]), and with respect to the type of object (i.e., target object, competitor object or one of two distractor objects).

Fourth, probability of fixation was computed for each type of object across time, operationalized as the percentage of gaze per participant per object type per time

[4]Since each trial was preceded by a fixation cross, participants tended to fixate the center of the screen in the beginning of a trial.

bin. The averages of these percentages across participants were used to plot the data (see Fig. 5.6 for an example from Dahan et al. 2000). The y-axis represents the proportions of fixations to the different types of objects and the x-axis the time in milliseconds. The values of the y-axis range from zero to one as the data is proportional. The zero point of the x-axis represents the onset of the noun. The vertical line in the plot corresponds to the offset of the noun.

5.4.4 Statistical Approaches

In VWP studies, participants' fixations on specific regions of interest are measured across the time course of an experimental trial. These regions of interest are defined on the basis of the different objects on the visual display. For example, in the case of a 4-object display, the monitor might be divided into a grid of four quadrants. On each trial each quadrant is classified as target, competitor or distractor. The typical question for the analyses is whether different regions of interest significantly differ in their likelihoods and timing of being looked at during different experimental conditions.

Currently, there is no consensus on the best way to inferentially test the observed differences. The difficulty with eye-tracking data is that fixations in any given region and at any given time are categorical (i.e. 0 or 1), whereas the independent measure, i.e. time, is continuous. A variety of analyses have been used to examine differences in proportion fixation (see special issue 59 of the *Journal of Memory and Language,* 2008). Traditional models compare proportional fixation to different objects over time using ANOVA or t-tests. The problem with such models is that they violate the underlying statistical assumptions of these tests (Barr 2008). Recently, researchers have therefore proposed more sophisticated statistical techniques such as multi-level logistic regression (Barr 2008), growth curve analyses (Mirman et al. 2008) and generalized additive models (Wood 2006).

Suggestions for Further Reading

Dussias, P.E., J.V. Kroff, and C. Gerfen. 2013a. Visual world eye-tracking. In *Research methods in second language psycholinguistics*, ed. J. Jegerski, and B. van Patten, 93–126. New York: Routledge.

Ferreira, F., M. Tanenhaus. 2008. Language-vision interaction [Special issue]. *Journal of Memory and Language*, 57(4).

Huettig, F., J. Rommers, and A.S. Meyer. 2011a. Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica* 137: 151–171.

References

Allopenna, P.D., J.S. Magnuson, and M.K. Tanenhaus. 1998. Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language* 38: 419–439.

Altmann, G.T.M. 2011. The mediation of eye movements by spoken language. In *The oxford handbook of eye movements*, ed. S.P. Liversedge, I.D. Gilchrist, and S. Everling, 979–1004. Oxford: Oxford University Press.

Altmann, G.T.M., and Y. Kamide. 1999. Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition* 73: 247–264.

Barr, D.J. 2008. Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of Memory and Language* 59: 457–474.

Boersma, P., D. Weenink. 2012. Praat: Doing phonetics by computer (Version 5.3.08) [Computer software]. http://www.praat.org/.

Clifton, C., A. Staub, and K. Rayner. 2007. Eye movements in reading words and sentences. In *Eye movements: A window on mind and brain*, ed. R.P.G. van Gompel, M.H. Fischer, W.S. Murray, and R.L. Hill, 341–372. Amsterdam: Elsevier.

Cooper, R.M. 1974. The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology* 6: 84–107.

Dahan, D., J.S. Magnuson, and M.K. Tanenhaus. 2001. Time course of frequency effects in spoken word recognition: Evidence from eye movements. *Cognitive Psychology* 42: 317–367.

Dahan, D., D. Swingley, M.K. Tanenhaus, and J.S. Magnuson. 2000. Linguistic gender and spoken-word recognition in French. *Journal of Memory and Language* 42: 465–480.

Duchowski, A.T. 2002. A breadth-first survey of eye-tracking applications. *Behavior Research Methods, Instruments, & Computers* 34(4): 455–470.

Dussias, P.E., J.V. Kroff, and C. Gerfen. 2013a. Visual world eye-tracking. In *Research methods in second language psycholinguistics*, ed. J. Jegerski, and B. VanPatten, 93–126. New York: Routledge.

Dussias, P.E., J.R. Valdés Kroff, R.E. Guzzardo Tamargo, and C. Gerfen. 2013b. When gender and looking go hand in hand: Grammatical gender processing in L2 Spanish. *Studies in Second Language Acquisition* 35: 353–387.

Ferreira, F., A. Foucart, and P.E. Engelhardt. 2013. Language processing in the visual world: Effects of preview, visual complexity, and prediction. *Journal of Memory and Language*, 69 (3), 165−182.

Grüter, T., C. Lew-Williams, and A. Fernald. 2012. Grammatical gender in L2: A production or a real-time processing problem? *Second Language Research* 28: 191–215.

Henderson, J.M., and F. Ferreira. 2004. Scene perception for psycholinguists. In *The interface of language, vision, and action*, ed. J.M. Henderson, and F. Ferreira, 1–58. New York: Psychology Press.

Hopp, H. 2013. Grammatical gender in adult L2 acquisition: Relations between lexical and syntactic variability. *Second Language Research* 29(1): 33–56.

Huettig, F., and J.M. McQueen. 2007. The tug of war between phonological, semantic and shape information in language-mediated visual search. *Journal of Memory and Language* 57(4): 460–482.

Huettig, F., J. Rommers, and A.S. Meyer. 2011b. Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica* 137: 151–171.

Lew-Williams, C., and A. Fernald. 2007. Young children learning Spanish make rapid use of grammatical gender in spoken word recognition. *Psychological Science* 33: 193–198.

Lew-Williams, C., and A. Fernald. 2010. Real-time processing of gender-marked articles by native and non-native Spanish speakers. *Journal of Memory and Language* 63: 447–464.

Loerts, H., M. Wieling, and M.S. Schmid. 2013. Neuter is not common in Dutch: Eye movements reveal asymmetrical gender processing. *Journal of Psycholinguistic Research* 42(6): 551–570.

Matin, E., K.C. Shao, and K.R. Boff. 1993. Saccadic overhead: Information-processing time with and without saccades. *Perception and Psychophysics* 53: 372–380.

Mirman, D., J.A. Dixon, and J.S. Magnuson. 2008. Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language* 59: 475–494.

McQueen, J.M., and M.C. Viebahn. 2007. Tracking recognition of spoken words by tracking looks to printed words. *Quarterly Journal of Experimental Psychology* 60: 661–671.

Oehler, H., and O.S. Gerretsen. 1973. *Duitse woordenschat—Alfabetische basisvocabulaire met systematische uitbreiding*. Groningen: Wolters Noordhoff.

Oostdijk, N. 2000. The spoken Dutch corpus. Overview and first evaluation. In Gavralidou, M., G. Carayannis, S. Markantonatou, S. Piperidis, G. Stainhaouer (eds.), *Proceedings of the second international conference on language resources and evaluation*, 887−893. Paris: ELRA.

Paris, G., A. Weber, and M.W. Crocker. 2006. *German morphosyntactic gender and lexical access*. Poster presented at the 12th annual conference on architectures and mechanisms for language processing (AMLaP 2006), Nijmegen, Netherlands.

Schneider, W., A. Eschman, and A. Zuccolotto. 2002. *E-Prime user's guide*. Pittsburgh: Psychology Software Tools Inc.

Snodgrass, J.G., and M. Vanderwart. 1980. A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory* 6: 174–215.

Spivey, M.J., and V. Marian. 1999. Cross talk between native and second languages: Partial activation of an irrelevant lexicon. *Psychological Science* 10: 281–284.

Tanenhaus, M.K., and J.C. Trueswell. 2006. Eye movements and spoken language comprehension. In *Handbook of psycholinguistics*, 2nd ed, ed. M. Traxler, and M. Gernsbacher, 86–900. Amsterdam: Elsevier.

Tanenhaus, M.K., M.J. Spivey-Knowlton, K.M. Eberhard, and J.C. Sedivy. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science* 268: 1632–1634.

Weber, A., G. Paris. 2004. *The origin of the linguistic gender effect in spoken-word recognition: Evidence from non-native listening*. Poster presented at the 26th Annual Meeting of the Cognitive Science Society (CogSci 2004), Chicago, IL.

Wood, S. 2006. *Generalized additive models: An introduction with R*. Boca Raton: Chapman & Hall/CRC Press.

Chapter 6
EEG and Event-Related Brain Potentials

Nienke Meulman, Bregtje J. Seton, Laurie A. Stowe and Monika S. Schmid

Abstract Event-related brain potentials (ERPs) have become a standard method in many areas of cognitive research, including second-language research, over the last decade and a half (Van Hell and Tokowicz 2010). ERPs can provide evidence which is central to the controversy on the similarity or difference of first and second-language processing. However, they provide a challenge which can be daunting for a large-scale multi-lab study, because there are so many technical details which vary from lab to lab, making it difficult to acquire data that is fully comparable across testing sites. In this chapter we will discuss a number of aspects of ERP measurement, focusing partly on experimental designs and partly on the way in which data from different languages and different labs can successfully be combined. These aspects will require somewhat more detail than the techniques treated in previous chapters.

Keywords Bilingualism · Grammatical gender · Second language acquisition · First language attrition · EEG · ERP · N400 · P600

6.1 ERPs and the Study of On-line Language Processing

6.1.1 Introduction to the Method

ERPs are used in language research to measure the online neural activity underlying language processing which unfolds while a comprehender reads or listens to linguistic material. They not only shed light on how language is processed in the brain, but can also reveal possible differences in underlying processing strategies between different groups of language users (e.g., monolinguals and bilinguals, or natives,

Electronic supplementary material The online version of this chapter (doi:10.1007/978-3-319-11529-0_6) contains supplementary material, which is available to authorized users.

M.S. Schmid et al., *Designing Research on Bilingual Development*,
SpringerBriefs in Linguistics, DOI 10.1007/978-3-319-11529-0_6

learners and attriters), thus offering the possibility of making qualitative inferences about the nature of these differences. This makes ERPs a powerful tool for studying online language processing in bilinguals.

ERPs are based on electroencephalography (EEG), a procedure that records real-time electrical activity in the brain through electrodes placed on the scalp. Since the EEG signal reflects the continuous, ongoing neural activity related to a multitude of brain processes which take place simultaneously (most of them unrelated to language), the brain response to a particular stimulus cannot be studied using the raw signal. This is different from the techniques discussed in the previous chapters: For example, measurements of eye-movements are strongly related to the task at hand, and each trial can start from a neutral baseline, where the gaze is relatively still and fixated on a central cross. The brain, however, is never ‘at rest’ in this manner, so the neural activity triggered by a certain linguistic stimulus has to be isolated from all other ongoing activity. In order to do this, many trials of the same type have to be recorded, all of which are time-locked to a certain kind of “event”—the presentation of a linguistic stimulus, for example, or the moment at which an ungrammaticality becomes evident. Through comparisons of the brain signal across these trials, brain activity which is unrelated in time to the event and thus occurs only sporadically and at different times throughout the stimulus will decrease or disappear, while the relevant event-related potential (i.e., the neural activity triggered by the type of stimulus used in the experiment) remains.

The temporal resolution of EEG measurements is very high (in the order of milliseconds), making it possible to accurately measure when a certain computational operation takes place in the brain.[1] There is one very obvious drawback to this method: Due to the fact that many stimuli of the same type have to be used, it is virtually impossible to obscure the purpose of a study by means of even larger numbers of distractors or fillers. This means that at the end of an EEG study, most participants will know which linguistic structure is being investigated.

ERP waveforms consist of a series of positive and negative voltage deflections. Predictable patterns of such brain activity typically encountered upon presentation of a certain type of stimulus (for example, a lexical or grammatical violation) are referred to as ERP ‘components’ (Luck and Kappenman 2012). Generally the names of ERP components are based on some combination of (1) their functional role (what type of stimulus modulates them, for example the Mismatch Negativity or Error Related Negativity), (2) timing (e.g., N400 which peaks around 400 ms after presentation of a word), (3) polarity (negative- or positive-going waveform as in N400, which is negative) and (4) scalp distribution (the region on the head where the component is encountered in the overall signal, e.g., Left Anterior Negativity). It is important to note that the scalp distribution does not directly indicate the location of the source of the neural activity; a left anterior negativity does not

[1]A full, in-depth discussion of the ERP technique is beyond the scope of this text; the reader is referred to Luck (2014) for a comprehensive introduction. We confine the discussion here to the points that are relevant for our purpose.

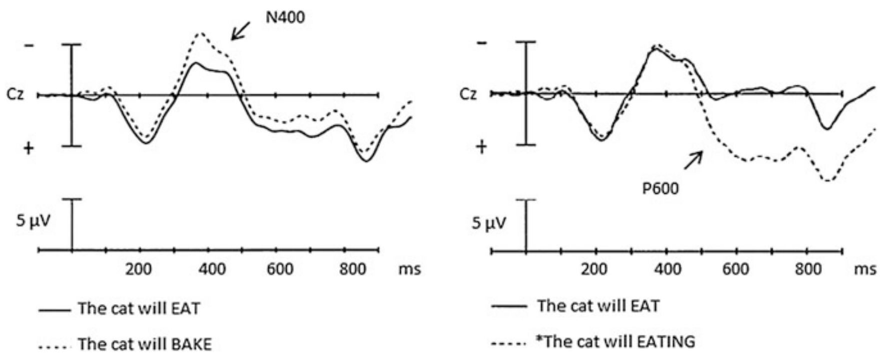

Fig. 6.1 Event-related brain potentials (ERPs) elicited by linguistically anomalous words encountered during sentence processing. Negative voltage is plotted up, and waveforms reflect measurement at electrode site Cz, which is located on the midline on the center of the scalp. *Left hand panel*: Semantically anomalous words ('The cat will bake the food …') elicit an N400 component with increased amplitude relative to the N400 elicited by the non-anomalous words ('The cat will eat the food …'). *Right-hand panel*: Syntactically anomalous words ('The cat will eating the food …') elicit a late positive shift (P600), relative to the ERPs elicited by the non-anomalous words. (Figures were adapted from Osterhout and Nicol (1999), their Fig. 1, p. 297 (detail) and are reprinted with kind permission from Taylor and Francis)

necessarily originate in the left anterior cortex, such as Broca's area. Scalp distribution is primarily important when it is argued that two effects which are otherwise comparable with respect to timing and polarity come from non-identical sources, because they have different scalp distributions within the same group of participants (e.g., Roehm et al. 2005). Researchers in the field of language processing in general and bilingualism in particular will be most interested in those components that are associated with language (see review by Swaab et al. 2012). The three best-known components of interest to linguistic research are the N400, the P600 and the Left Anterior Negativity (LAN).

The N400 is a negative peak in the signal which occurs approximately 400 ms post stimulus (see Fig. 6.1) with a posterior (back of the head) and bilateral (on both sides of the head) distribution. It has been shown to be modulated by semantic factors (Kutas and Hillyard 1980), with a larger amplitude when a word is encountered that is perceived to be unexpected within the context of the utterance. The left-hand panel in Fig. 6.1 illustrates modulations of the N400 in a comparatively naturalistic utterance, such as 'The cat will *eat* the food I left on the porch' (solid line), as opposed to an unexpected construction, e.g., 'The cat will *bake* the food I left on the porch' (dotted line), where the unexpected word *bake* modulates the ERP component. (Italics will be used throughout this chapter to indicate the target word, i.e., the word to which the response is time-locked in the ERP signal.)

The amplitude of the N400 has been shown to depend on the expectancy of the upcoming word in a sentence: Words that are semantically unexpected elicit larger amplitude N400 responses than words that are more expected given the preceding sentence and discourse context (Kutas et al. 1984), as well as words that are more

closely related to the previous context (Federmeier and Kutas 1999) and higher frequency words (Van Petten and Kutas 1990). Therefore, the N400 is likely to reflect semantic processes of lexical access and/or integration (Brouwer et al. 2012; Friederici et al. 1993; Holcomb and Neville 1991).

The second ERP component that has been found reliably in relation to language processing is the P600. It differs from the N400 in polarity, timing and (to a lesser extent) scalp distribution and is found in response to the processing of morphological and other syntactic information (Molinaro et al. 2011; Osterhout and Holcomb 1992). It is a late positivity (usually detectable in the 500–800 ms time window), broadly distributed over posterior sites. This component occurs in response to morphological and syntactic errors, such as violations of subject-verb agreement and case inflections (see Molinaro et al. 2011, for an overview), and with garden pathing in ambiguous sentences (Osterhout et al. 1994). The waveform in the right-hand panel of Fig. 6.1 shows the response to sentences of the type 'The cat will *eating* the food I left on the porch'. It is often believed that the P600 reflects a controlled mechanism associated with reanalysis and repair which is triggered when incoming words cannot be incorporated into the syntactic structure that was initially built (Friederici et al. 1996). However, there is considerable debate over the precise functional interpretation of the P600 (Brouwer et al. 2012; Coulson et al. 1998; Kaan et al. 2000). Although it is important to be aware of this debate for the purpose of second-language research, the P600 can be used as a sign of native-like processing in highly proficient populations of learners without necessarily subscribing to any given interpretation.

A last component which has sometimes (but less reliably) been found in response to violations of morphosyntactic rules is the left anterior negativity (LAN, see Fig. 6.2). The LAN is typically observed at around 300–500 ms after a violation has been detected and often occurs in conjunction with the P600 in a biphasic pattern (Friederici 1995; Molinaro et al. 2011). This component has also been associated with increased working memory load in *wh*-questions (Coulson et al. 1998; King and Kutas 1995; Kluender and Kutas 1993). While the N400 and the P600 have been found reliably and consistently in response to semantic and morphosyntactic violations, respectively, across a range of languages and native and non-native populations, the LAN is more elusive, and it remains unclear what exactly determines whether or not it will occur in any given investigation (for an overview, see Molinaro et al. 2011).

These are the three most commonly found ERP components linked to language processing.

Knowledge about these components generated by studies with native speakers can inform us about how best to design an experiment dealing with bilinguals. Luck (2005) suggests focusing a given experiment on only one or two components, preferably well-established, large and reliable ones, and using well-studied experimental manipulations to avoid ambiguities in interpreting ERP effects. This is particularly important in L2 research, where so many factors can affect the outcome. For this reason, we are here focusing on components which are relevant to sentence processing. There are several others, for example the Mismatch Negativity (MMN) which

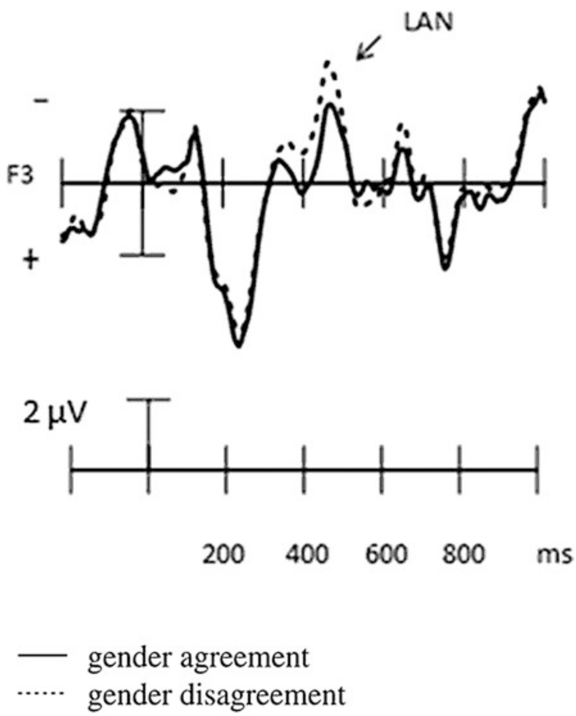

Fig. 6.2 A left anterior negativity (LAN) associated with gender mismatch in German. F3 is an electrode located to the *left* of midline and over frontal areas of the scalp. Figure adapted from Gunter et al. (2000), their Fig. 1, p. 561 (detail) and reprinted with kind permission from MIT Press

has been successfully used to investigate the development of phonological processing in L1 (e.g., Kuhl 2010) and L2 (e.g., Peltola et al. 2003), as well as the focus positivity (Dimitrova et al. 2012) and the Closure Positive Shift (Steinhauer et al. 1999) which reflect aspects of prosodic processing. Likewise, ERPs in response to pictures which are to be named in a designated language show a negativity which is thought to be related to the inhibition of the undesired language (Misra et al. 2012).

ERPs have become one of the most frequently used methods in neurocognitive research because they reflect the fine-grained on-line response to various inputs on the timescale in which linguistic processing takes place—in contrast to other neuroimaging methods, such as fMRI, which has excellent spatial but poor temporal resolution. For research on language processing, ERPs have the advantage that the participant does not have to carry out any task except normal comprehension and are therefore assumed to reflect naturalistic processing better than, for example, grammaticality judgment tasks, which involve metalinguistic skills. This allows the researcher to ask detailed questions about the ways in which populations differ from each other and in which they are the same.

At the same time ERPs rely on many responses of the same kind, which means that a large number of trials is necessary in order to obtain a clean signal from each individual participant. Similarly, individuals may differ substantially from each other in terms of the size of the response and even the scalp distribution of a given response, and a large sample size is therefore necessary to accurately represent any

given population. For both of these reasons, data collection using this method represents a considerable investment in terms of time, for the participant but particularly for the researcher. For bilingualism research, this is further exacerbated by the number of other factors which may affect the results (see Chap. 2). In conclusion, the method is not for the faint of heart, but we are convinced that, if the experiment is properly set up, the potential gains outweigh the difficulties.

6.1.2 Monolingual and Bilingual Processing

The ERP components sketched above were initially discovered in monolingual native populations. More recently, they have also been extensively studied among second-language learners and bilinguals. In this context, it has been found that even relatively low-proficiency speakers quickly show N400 responses to semantically unexpected items in a manner similar to processing among monolingual natives (McLaughlin et al. 2004), although sometimes with a smaller amplitude or delay in onset (Ojima et al. 2005; Weber-Fox and Neville 2001). In contrast, factors such as proficiency, age of onset of L2 acquisition and cross-linguistic similarities between L1 and L2 substantially modulate P600 and LAN effects in morphosyntactic processing (Foucart and Frenck-Mestre 2011; Meulman et al. 2014; Sabourin and Stowe 2008; Steinhauer et al. 2009; Weber-Fox and Neville 1996). As with the N400, differences between L2 learners and natives for these components may be visible in a decreased amplitude and/or delayed onset of the component. However, there are also cases where even advanced L2 learners show completely different ERP components from those observed among natives in response to a particular structure or violation, or where the processing of the violation does not differ at all from that of the target-like structure.

For example, some studies have found an N400 instead of a P600 in early learners for some grammatical violations (McLaughlin et al. 2010; Weber and Lavric 2008). It has been suggested that this may be the outcome of a lexical (instead of a syntactic) strategy during processing, due to limited overall proficiency (Clahsen and Felser 2006; McLaughlin et al. 2010; Steinhauer et al. 2009). The LAN, which is elusive even among native speakers, has rarely been found in late L2 populations; the few exceptions involve either early learners (Weber-Fox and Neville 1996) or regular constructions which are very similar between the L1 and the L2 involved (Rossi et al. 2006).

Generally speaking, the differences between L2 learners and natives are larger at lower L2 proficiency levels, for learners with a later age of onset of acquisition, and for L2 learners whose first language is typologically further removed from the target language with respect to the construction under investigation. Whether second-language speakers, in particular those who begin acquiring the L2 after puberty, can show fully native-like neural responses to linguistic input, and in particular whether this ability is achievable across the board for *all* grammatical structures, is a topic that is still under debate (e.g., Steinhauer et al. 2009).

6.1.3 ERPs and Grammatical Gender

The multi-lab project discussed throughout this volume focused on gender concord, which is among the most intensively studied topics in the debate on bilingual processing. For this grammatical structure, the differences between L1 and L2 are particularly multi-faceted. They include differences at the level of the presence or absence of grammatical gender, the specifics of the grammatical agreement rules (on which sentence constituents is gender concord marked?) and differences at the level of lexical classification (a particular item may be masculine in one language but feminine in another). For these reasons, gender concord appears to be one of the most difficult features to acquire for second-language learners.

Studies of native populations have very reliably found a P600 response to violations of gender agreement across a range of languages (Spanish: Barber and Carreiras 2005; Italian: Molinaro et al. 2008; French: Foucart 2008; German: Gunter et al. 2000; Dutch: Hagoort and Brown 1999; Loerts et al. 2013; Sabourin and Stowe 2008; Van Berkum et al. 1999). In some cases, a LAN has also been observed (German: Gunter et al. 2000; Spanish: Barber and Carreiras 2005) but not in others (Dutch: Sabourin and Stowe 2008; French: Frenck-Mestre et al. 2009).

L2 learners show an even more complex pattern. Previous findings suggest that the LAN is unlikely to be found in late learners; unfortunately there are to date no studies of gender concord among early L2 learners. Even the P600, which is frequently reported even for late learners with respect to other grammatical features, is highly variable. Some learners show N400 effects in response to article-noun combinations which violate gender agreement, for example when the gender of the target noun is not the same as in their native language or when the agreement context is relatively infrequent (Foucart and Frenck-Mestre 2012). In other studies, a P600 is found, but this component may diminish or even disappear when the type of agreement changes from adjacent elements to longer distance agreement or when the agreement rules differ between L1 and L2 (Foucart and Frenck-Mestre 2011; Frenck-Mestre et al. 2009; Meulman et al. 2014; Sabourin and Stowe 2008).

There are two striking gaps in the literature. First, there are very few investigations of the impact of age of acquisition on L2 processing (the only exception being Weber-Fox and Neville 1996); most studies involve learners whose L2 acquisition started post-puberty, in their teens or twenties. Second, to date investigations of grammatical processing among bilinguals in general and gender concord processing in particular have focused exclusively on the L2 of these speakers, and on the issue of the attainability of native-like processing. The question of whether extended periods of L2 use, language dominance reversal and L1 attrition can also impact on first language processing as measured by the EEG, and whether the brain responses can move away from those observed in monolingual natives and towards the patterns typical for L2 learners, has not previously been addressed by means of the ERP technique.

6.2 Designing an ERP Experiment

6.2.1 General Design Issues

There is a wealth of existing knowledge about ERP components which demonstrates that ERPs can be affected by a range of linguistic factors (characteristics of the stimulus). It is therefore vital to control for these factors in designing an experiment in order to exclude effects that are not due to the target manipulation (e.g., a violation of agreement on a particular morphosyntactic feature) but to some other aspect of the material. Among the factors that are important in this context are plausibility, frequency, complexity, sentence length and linear position of the violation as well as other visual or auditory characteristics of the stimulus. Each of these factors may quite conceivably have a different or stronger effect for populations which are not monolingual.

It is a well-established finding that the ***plausibility*** of a sentence modifies the amplitude of the N400 as it affects the ease with which new words can be integrated into the existing sentence frame. It is therefore important to ensure that all sentences are rated as more or less equally plausible by native speakers. For bilingual populations the ratings should ideally also be conducted by members of the target L2 population (note that individual speakers who have given such plausibility ratings in the design phase should not be included as participants in the actual experiment): Sentence interpretation is based on real world knowledge (Hagoort et al. 2004) and this may be affected by cultural differences. Secondly, ***word frequency*** also affects the amplitude of the N400, as mentioned above. Very infrequent words can furthermore elicit a late positive component which occurs in the same time window as the P600, and, of course, higher proportions of infrequent words increase the likelihood that the L2 populations may not know some of these items. It is therefore important that the frequency of all target items be assessed, preferably on the basis of a large database such as CELEX (Baayen et al. 1995) or the English and Dutch Lexicon projects (Balota et al. 2007; Keuleers et al. 2010). The frequency assessment is of particular importance for the target words within the phrase where the violation occurs. However, as familiarity may affect the overall plausibility or parseability of a sentence, the frequency of the other items should also be kept at the same (high) level. Again, investigations of bilinguals present an additional challenge, since it is necessary to ensure with reasonable certainty that all lexical items are familiar to the participant population. This can be done by means of a familiarity pretest (to be conducted together with the plausibility assessment).

Sentence length and ***complexity*** affect slow waves[2] (Phillips et al. 2005; Van Petten and Kutas 1991), as well as the P600 (Kaan et al. 2000). Again, these factors are particularly important in the context of second-language learners, who might

[2]Slow waves, also called DC potentials, begin to emerge at least 500 ms after stimulus onset and often last for several seconds. They usually appear as negative deflections and have a less distinct peak.

encounter more difficulty processing complex and longer sentences than monolingual natives, and in this manner the effects which are being studied can be obscured or diminished. This might lead to differences between the experimental and the control population which could then be falsely ascribed to the target grammatical feature (i.e., a false positive). Lastly, it has been suggested that responses to items in ***sentence-final position*** may differ from those to items placed elsewhere in the sentence (Osterhout 1997), and so this position should be avoided for the target violation where possible.

Keeping all of these potentially confounding factors constant across the experiment is no small challenge. A powerful approach to isolating effects that are caused by a single factor is to calculate ***difference waves*** between sentence pairs that, other than for the target manipulation, are identical to each other in terms of their phonological, syntactic and semantic-pragmatic features. In other words, the same sentence is presented twice, once in a grammatically correct form, and once with the target violation (as illustrated in the example in Fig. 6.1). The EEG signals in response to both versions can then be subtracted from each other, resulting in a difference wave (Fig. 6.3) which contains the components affected by the manipulation of the stimulus type without the distraction of all the other electrophysiological changes elicited by both types of stimuli.

While this type of within-item design allows the researcher to zoom in on the variance elicited by the violation, it is problematic for other reasons: ERPs are sensitive to ***repetition***, including the repetition of a sentence context (Besson et al. 1992). The same participant should therefore never be exposed to both the grammatical and the ungrammatical version of the same sentence. The solution is to create such correct/incorrect pairs for all stimuli but distribute them across different

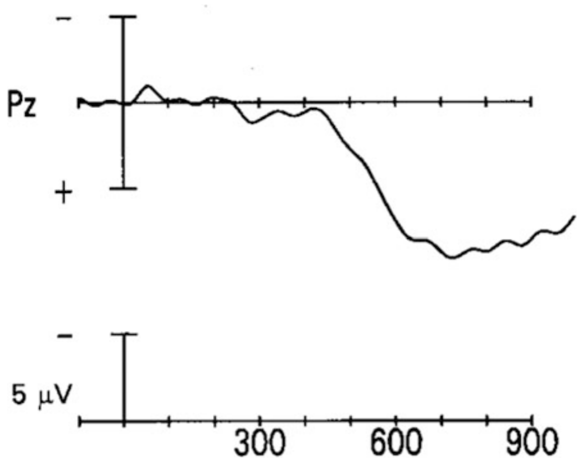

Fig. 6.3 Difference wave 'The cats will eating …' minus 'The cats will eat …'; this demonstrates the time and size of the difference more clearly than the overlaid waveforms and is a useful extra visualization of the data. Figure adapted from Osterhout and Nicol (1999), their Fig. 7, p. 303 (detail) and reprinted with kind permission from Taylor & Francis

lists, so that each participant will only encounter either the grammatical or the ungrammatical version.

Given the complexity of all of the factors named above, it is advisable that the investigation should not focus on only one grammatical feature. In particular where features are concerned that are assumed to be difficult to acquire, it is wise to also include a feature which L2 learners can be assumed to master at a relatively low level of proficiency. In this manner, it can be assessed whether the overall experiment is set up correctly and whether the bilingual participants respond as expected to grammatical violations that they do master. Such a test feature can also serve as a complement to the more off-line measures of general proficiency described in Chap. 2.

Investigating multiple grammatical structures within the same experiment, however, may inflate the total number of stimuli beyond the limits of feasibility. The ***number of trials*** necessary to obtain a reliable ERP depends on a number of factors, the most important being the "signal-to-noise ratio", that is, the size of the signal (the ERP) relative to the size of the noise (the background EEG), and this will vary across setups, experiments and participants. In language experiments, more than 20 stimulus presentations per condition per participant are commonly regarded as necessary to obtain a good clean average ERP, although very strong and reliable effects, such as the P600 in response to a syntactic violation, may elicit an effect with fewer items. However, limiting the number of stimuli accordingly may mean that other components that are also present but not as strong, such as a LAN, remain undetected. Unfortunately, presentation of a large number of sentences containing the same grammatical manipulation will almost inevitably cause participants to become aware of the manipulation. This may lead to two things: First, participants may pay conscious attention to the target violation, or conversely they may begin to habituate to it. To some extent these effects can be minimized by the inclusion of filler items. A large number of items and fillers, however, may lead to an experiment that is very long, particularly if many factors are tested. Good design therefore often requires compromise between power and coverage, and it may not be possible to test all of the constructions or variants in which the researcher is interested within one single experiment. As a rule of thumb, in order to avoid both fatigue and frustration among participants (both of which can lead to problematic results, in addition to being ethically undesirable), an EEG experiment should include a number of breaks and not last longer than 1½ h.

The last issue to be discussed in the context of experimental design concerns the ***mode of presentation***. Timing is a critical factor in ERP research; it is necessary to know with absolute precision the moment at which a grammatical violation becomes detectable in order to time-lock the EEG analysis to that moment. For this reason, most studies opt for written stimuli and Rapid Serial Visual Presentation (RSVP), that is, showing sentences one word at a time for a brief period (typically around 250 ms) in the middle of the computer screen. However, this type of presentation, while convenient, may seriously disadvantage bilingual participants and thus lead to a confound between native and non-native populations. First, quickly and automatically detecting a violation in written text depends heavily on the level of reading experience, which will differ between populations, since the

L2ers will have begun reading their second language much later and may not read it as frequently. Second, RSVP places more demands on working memory than does normal reading, and again this factor may impact native versus L2 populations differently (see Chap. 2). Lastly, processing of the written form may be influenced by the familiarity with the L2 script, producing another source of differences between some L2 learners (whose L1 uses a different script than the L2) and monolinguals (for a direct comparison of visual and auditory stimuli in an ERP study of bilingual processing see Meulman et al. 2014). Although constructing auditory stimuli and embedding them in the experiment is more challenging and time-consuming than creating a visual experiment, we feel that the benefits of using the spoken mode outweigh the disadvantages when it comes to investigations of bilingual processing.

6.2.2 Multifactorial Considerations

In this section we will outline the steps that are necessary when designing and setting up an EEG experiment of bilingual processing, based on the multi-population and multi-lab investigation of grammatical processing among bilinguals on which the present volume is based. Like the eye-tracking study described in Chap. 5, the ERP study focused on grammatical gender. The EEG experiment was different from the others described in the present volume in that the gender concord condition was augmented by a second grammatical feature, namely agreement between a non-finite main verb and its verbal auxiliary. This feature is assumed to be acquired and retained more easily and at lower proficiency levels than grammatical gender (see Loerts 2012; Sabourin and Stowe 2008), and the non-finite verb condition was thus included to serve as a proficiency control for the gender agreement condition in the manner suggested above.

The main interest of the current study was to assess under what circumstances bilingual populations differ at the neuronal level with respect to the processing of their second language (L2 learners) or their first language (L1 attriters) from predominantly monolingual speakers. Factors which are expected to influence the outcome are the age at which the individual became bilingual, the speaker's proficiency in the language under investigation and the amount of use which is made of the L1 and the L2 respectively. For the L2 learners the presence of a similar construction in the L1 is also relevant. Across the two languages under observation (Dutch and German) a further factor relates to the salience and reliability of gender agreement in the language being tested. As was pointed out in Chap. 1, the two experimental languages differ considerably with respect to this point. If salience and reliability affect acquisition and/or attrition, this should be reflected in participants' brain responses.

Target populations included a variety of L2 Dutch and German speakers both from languages which mark gender concord morphologically and from languages which do not, as well as Dutch and German attriters (see Chap. 1). Acquiring data

for these populations necessarily involved EEG labs in several cities spread over four different countries (see Chap. 3). We will discuss the combination of these data in more detail in Sect. 6.5.

6.3 Materials

The ***non-finite verb control condition*** is virtually identical in German and in Dutch, reducing the difficulties of creating matched materials across the two languages. In both languages, periphrastic verb constructions with a finite auxiliary verb require the past participle form of the main lexical verb, while constructions with a finite modal auxiliary require the infinitival form. Forty-eight pairs of sentences were constructed, and in one member of each pair the correct past participle or infinitive was replaced by the ungrammatical infinitive or past participle verb form (see Table 6.1). It should be noted that the type of construction (infinitive vs. past participle target form) was not included as a factor in the analyses but merely served to decrease the predictability of the construction in order to prevent participants habituating to the violation.

Constructing stimuli for the ***gender condition*** was somewhat more complex, due to the crosslinguistic differences. Both Dutch and German mark grammatical gender on various aspects within and across noun phrases (determiners, adjectives, quantifiers, pronouns etc.) and in both languages gender agreement interacts with definiteness (see Chap. 1). An investigation of all of these different agreement types would have been beyond the feasibility of a single study, given the comparatively large number of tokens necessary for any particular condition in an EEG experiment. We therefore decided to limit the grammatical gender agreement manipulations to two constructions: (1) definite determiner–noun (D_{gen}–N) structures, in which gender agreement is expressed on adjacent elements in both languages, and (2) definite determiner–adjective–noun (D_{gen}–Adj–N) structures, in which gender agreement is non-adjacent. The inclusion of the latter has two reasons; first, agreement across a longer distance has been shown to affect the P600 effect in L2 populations (Foucart and Frenck-Mestre 2012), and second, the presence of an intervening element prevents the use of a chunking strategy to mimic gender knowledge.

The design and materials of the Dutch version of this experiment were partly based on the study of second-language acquisition of Dutch gender reported by Loerts (2012).[3] The materials consisted of 96 pairs of sentences. For each grammatical sentence, an ungrammatical version was constructed which contained a

[3]Loerts' study includes a third type of sentence, namely indefinite determiner—adjective—noun, and as a result had less tokens per condition. In order to ensure validity and reliability for our study, which (unlike Loerts' investigation) included experiments from different populations and data acquired at different sites, we decided to test only the two structures mentioned and increase the number of tokens per structure.

Table 6.1 Samples of the experimental materials per condition of the Dutch and German version of the example experiment, showing the possible article–noun and non-finite verb combinations that created the grammatical and violation conditions

Condition	Dutch example sentence	German example sentence	Number of items per list
Gender, D–N	Na het ongeluk werd de vrouw per ambulance naar het/*de ziekenhuis gebracht	Nach der Schlägerei ist das/*der Auge des Angestellten von der Krankenschwester versorgt worden	24/24
	After the accident the woman was brought to the$_{neut}$*/*the*$_{com}$ *hospital by ambulance*	*After the fight the*$_{neut}$*/*the*$_{masc}$ *eye of the worker was treated by the nurse*	
Gender, D–Adj–N	Het verkeer raast elke ochtend over de/*het lange weg die de mensen naar Rotterdam brengt	Glücklicherweise ist der/*das frühe Beginn der Konferenz nach hinten verschoben worden	24/24
	Traffic rushes by every morning along the$_{com}$*/*the*$_{neut}$ *long road which takes people to Rotterdam*	*Fortunately the*$_{masc}$*/*the*$_{neut}$ *early beginning of the conference was postponed*	
Non-finite verb	Ze hebben van de directeur een cadeautje gekregen/*kregen voor de kerstdagen	Leider hat die Rose diesen Herbst noch nicht geblüht/*blühen, obwohl wir sie jeden Tag gegossen haben	24/24
	*From the manager they have received/*receive a present for Christmas*	*Unfortunately this fall the Rose has not yet bloomed/*bloom, although we watered it every day*	

gender violation on the article (see Table 6.1), leading to a total of 192 sentences. The pairs were distributed evenly over the two genders of Dutch (48 sentences contained a common gender noun, and 48 sentences a neuter one), and over the two types of constructions (D-N and D-Adj-N), so that the final set of grammatical sentences comprised 24 D-N sentences with common and 24 D-N sentences with neuter nouns, and the same number of D-Adj-N sentences. In creating the set of German materials we attempted to emulate as closely as possible the same manipulations of gender. The gender contrasts in this language were therefore limited to masculine and neuter nouns, in order to avoid contrasting the Dutch two-way distinction with a three-way one in German (the feminine was excluded due to the form overlap of the determiner for all genders in the plural).

It should be noted here that the structure of the sentences between the two languages differed somewhat, due to cross-linguistic differences. In the Dutch experiment, the manipulations were usually encoded on an object or a prepositional phrase to allow for the target noun to appear towards the middle of the sentence. For German, using the target in the object position would have implied using oblique case forms in which unambiguous gender identification is not always possible (since, for example, masculine and neuter are identical in the dative, while the nominative and accusative forms are identical for the neuter but not for the masculine). We therefore decided to

use items that were consistently in the nominative case, where gender is most unambiguously identifiable. However, inanimate items can plausibly be used as the subject-agent of a sentence with only a limited set of verbs. We therefore decided to frame the German stimulus sentences in the passive. The passive sentences included sentence-initial topicalized elements to induce inversion of the subject to the position after the finite verb in which it is unambiguously nominative. This strategy had the added advantage that the sentences were structurally more similar to the Dutch ones, with the target item always appearing in post-verbal position. Example items of each condition are given in Table 6.1 and a full list of all experimental materials can be found in the online supplementary material.

Following Loerts (2012), a third factor was taken into account for the Dutch experiment. This factor relates to the predictability of the upcoming target noun, a factor also known as cloze probability (Kutas et al. 1984). The Dutch experiment included this experimental manipulation in the gender conditions in order to test effect of lexical constraint, which has been found to affect native speakers (Gunter et al. 2000; Loerts et al. 2013) and might be similarly important for learners and attriters. Two versions of each of the sentence pairs were therefore created, one with a highly constraining context, and one in which the target noun was unpredictable from the preceding information (see examples in Table 6.2). Hence, the Dutch experiment in fact contained quartets instead of pairs of sentences.

If treated as a full factorial design, the number of items for each condition is further halved to 12. As was pointed out above, it is desirable to keep the number of stimuli per condition above 20, in order to avoid a poor signal-to-noise ratio (see Sect. 6.2.1), particularly for between-group designs. Increasing our stimuli to this number, however, would have inflated the total number of items, and thus the length of the experiment, beyond the limits of feasibility. We therefore opted to initially test the main effect of cloze probability (without differentiating between the D-N and the D–Adj–N structures) between populations. Unfortunately, the structural differences between the Dutch and German sentences made it impossible to create high cloze sentences in the German experiment, since the topicalized phrase and the passive auxiliary *sein* 'to be' which preceded the target nouns provided insufficient constraint. This was deemed acceptable, as the information from this

Table 6.2 Example sentences of the cloze probability manipulation in the grammatical gender sentences of the Dutch experiment

Condition	Dutch example sentence	Number of items per list
High cloze	Johan betaalde de motor met het/*de geld dat hij voor zijn verjaardag kreeg	24/24
	*Johan paid for the motorcycle with the$_{neut}$/*the$_{com}$ money that he received for his birthday*	
Low cloze	Johan was erg blij met het/*de geld dat hij voor zijn verjaardag kreeg	24/24
	*Johan was very happy with the$_{neut}$/*the$_{com}$ money that he received for his birthday*	

manipulation would be available for the Dutch attriters and learners and no difference is expected between languages for this aspect of processing, unlike for gender agreement.

The ***frequency*** of the target items was assessed on the basis of the CELEX corpus (Baayen et al. 1995) for Dutch and the DeReKo corpus (Institut für Deutsche Sprache 2010) for German. All target items fell within the highest frequency range in both languages, and should therefore be familiar to bilinguals as well as monolinguals. ***Plausibility*** and ***comprehensibility*** of experimental sentences was determined by means of pre-testing. Such a pre-test can either be done in the form of a formal written questionnaire (e.g., the grammatical variant of all stimulus sentences being rated for plausibility on a 5-point scale) or in an informal manner during which the researcher presents each individual with the stimulus sentences and asks them to comment on plausibility and comprehensibility. The latter approach was taken in the current study, as auditory presentation was considered important as well as the possibility to receive qualitative feedback. Both natives and second-language learners of each target language participated in these pretests, and in those cases where there were any problems with plausibility or comprehensibility, the stimuli sentences were adjusted.

Cloze probability was pre-tested by presenting native speakers of Dutch with incomplete written versions of each sentence and asking them to fill in the word they thought would follow after the determiner or the determiner and adjective. This allowed us to assess whether items were highly expected or less expected. Since in the German experiment, the items that were used were all unpredictable, this pre-test was not conducted for the materials in this language.

We suggested above that auditory presentation is a preferable choice when testing participants who may have limited experience reading a given language. This is likely to be especially true for attriters with an early age of emigration and late second-language learners who arrive in the new country as adults (some members of both populations may even be illiterate in the language under investigation), so auditory presentation was employed for this experiment. All sentences were recorded by a female native speaker of Dutch and German, respectively. Both speakers spoke the standard variety of the language and had considerable elocution training and experience. They received practice in producing the correct and incorrect sentences with their best approximation of normal intonation, and each sentence was recorded three times in both its grammatical and ungrammatical version. The clearest versions were chosen.

When recording ungrammatical sentences, it should always be kept in mind that, no matter how carefully the speakers are trained beforehand, it is possible that unconscious cues to the ungrammatical target remain present in pronunciation, hesitations or intonation (evidence for prosodic cues to unexpected material has been found by Dimitrova et al. 2009; Dimitrova 2012, shows that prosodic patterns can be noticeably initiated before word onset). Therefore, for each sentence, two versions of the recording were constructed: one consisting of the unaltered original (grammatical or ungrammatical recording), while in the other the target region was spliced together with the context preceding the target word from the other variant.

This resulted in four versions of each sentence: grammatical and ungrammatical version with and without splicing. Noise reduction and volume normalization were also applied to all sound files to reduce variability.

The resulting stimuli were divided across lists based on a Latin Square design,[4] to ensure that no participant would encounter different versions of the same sentence, since repetition can influence ERP responses (see above). For both languages, the factor splicing (spliced vs. unspliced) was treated as a counterbalancing variable rather than as part of the design. The Dutch experiment also included cloze probability (high vs. low), resulting in a total of eight ($2 \times 2 \times 2$) sentences containing each target noun and therefore eight lists, while there were only four (2×2) for the German experiment. In addition to the experimental sentences ($n = 144$), all lists contained a number of well-formed filler sentences, lowering the overall proportion of incorrect sentences to around a quarter, which makes the task more similar to natural language processing, and also reduces the noticeability of the target violation (see above). The proportion of violations in an ERP experiment is an important factor to keep in mind, since probability and saliency of manipulations have been shown to affect the size of ERP components (Coulson et al. 1998).

6.4 Experimental Procedure

At most of the testing sites (see Chap. 3) the experiment was carried out in an electrically shielded and sound attenuated chamber. In locations where no such chamber existed, it was necessary to ensure that surrounding noise and electrical influences were decreased to a minimum. Participants were comfortably seated in front of a computer screen, at about 60 cm distance. Visual stimuli, consisting of a fixation cross and the grammaticality judgment question, were presented on this screen. Loudspeakers were placed to the left and right side of the screen and the volume was adjusted to be clearly audible, but not too loud, before the experiment started. After each sentence was completed, the participant was prompted to make a grammaticality judgment by means of a button press on the keyboard. The presentation of the auditory sentences and the recording of accuracy and reaction times of the grammaticality judgment was done by means of E-prime (Schneider et al. 2002a, b).

Participants were asked to avoid moving any parts of their body and to not make eye-movements or blink during sentence presentation, using the fixation cross to help maintain a constant gaze. This is important as muscular activity, blinks and eye movements can distort the signal measured by the scalp electrodes (see below on

[4] A Latin Square design is a way of distributing items across subject lists, resulting in one version of a given sentence per list, and enough lists so that each version appears on one (and only one) list. The stimuli given to each participant are based on one of the lists, ensuring that s/he will see only one version of each experimental sentence.

how to deal with the artifacts that cannot be avoided). Before the experiment began, there was a practice block with materials similar to those tested in the experiment, to allow the participants to get used to the situation and task, and if necessary to ask questions. The experiment itself was broken up into four blocks to allow the participant to move freely and relax during breaks; the duration of the breaks between blocks was determined by the participant. Depending on the length of the breaks, the experiment lasted about one hour.

6.5 Data Recording and Analysis

One of the biggest challenges of an ERP investigation which compares data acquired at different testing sites is that there are a number of differences between EEG recording setups which can have important consequences for the resulting raw data. Some researchers may be able to bring their own portable EEG setup to each site, and can thus avoid these difficulties. Where this is not an option, a compromise needs to be made between locations with similar setups and the central criterion of access to sufficient numbers of the target population. Fortunately, in many cases the data can be processed in such a way that it will be comparable across locations, as in the present study.

The most important characteristics of the set-up depend on the amplifier and the caps employed, in combination with the registration software. A number of important factors depend on these two pieces of hardware: Firstly, EEG systems differ with respect to how many electrodes are measured and where they are located on the scalp. Secondly, there are differences in how eye-movements and blinks are registered. A third important difference contains the on-line reference electrode and its position. Lastly, there are differences in the sampling rate (how many measurements are taken per second) and the extent of electrical noise and in the signal. We discuss each of these factors briefly below. For more extensive discussion of the consequences of each of these parameters we refer the reader to Luck (2014).

Number and placement of electrodes depend jointly on the amplifier (how many channels are available?) and the cap (how many electrodes can it contain?). The set-up with the smallest number of electrodes available across all testing sites determines how many electrodes can be used in the final analysis, so amplifiers that are equipped to deal with a larger number are to be preferred. Thirty-two are usually considered acceptable for investigations of N400, P600 or LAN effects, as well as for most other relevant language responses. The positions may differ slightly from site to site as well, depending on the caps which are used locally, but the majority of labs use positions based on the international extended 10–20 system (see Fig. 6.4). Caps with smaller numbers of electrodes tend to use the original sites to ensure a wide coverage of the head, so there is likely to be overlap between the sites for most if not all electrodes. A common procedure in the analysis phase is not to use the signal acquired from every individual electrode, but to establish so-called Regions of Interest (ROIs). The signal from the electrodes located in this region is then

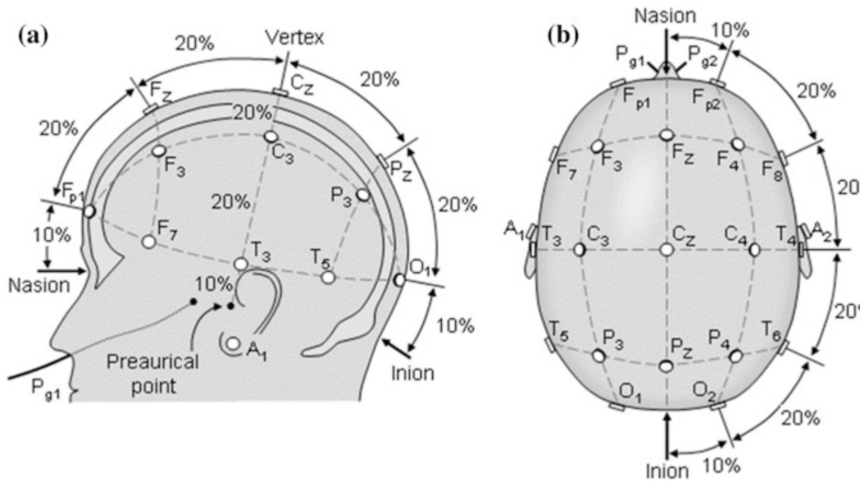

Fig. 6.4 The international 10–20 system uses four reference points (nasion, between nose and brow; inion, at the outer edge of the tentorium; and the joint in front of both ears to place electrodes at regular intervals based on percentage increments. Figure from Malmivuo and Plonsey 1995; redrawn from Sharbrough et al. 1991)

averaged together. This common practice is also very useful when dealing with caps and amplifiers that support different overall numbers of electrodes.

Blinks and eye movements cause distortions in the EEG signal in that they cause waveforms of far greater amplitude than the signal measured from the brain (see Fig. 6.5). For this reason they need to be removed or corrected (Gratton and Coles 1989). In order to detect the muscular movements associated with these two types of activity, electrodes placed near the eye are used to register the occurrence of blinks and vertical or horizontal eye movements (electro-oculogram or EOG). Set-ups can consist of, for example, single electrodes beneath one eye and next to it, to register vertical and horizontal eye movements respectively, or of bipolar electrode pairs, placed above and below the pupil and at both sides. The difference between these electrodes provides a more accurate measure of the eye movements and blinks. Encouraging the participants to fixate on a single point diminishes the number of saccades, and asking people to refrain from blinking during sentences decreases the number of blink artefacts. Blinks can create an additional problem for experiments using written presentation by the RSVP method, as the reader may miss essential information during a blink, which typically lasts about 200 ms.

Whether to reject trials that are contaminated by blinks or eye movements or whether to try to correct for them by reconstructing the original signal is an unresolved issue. Rejection leads to data loss, but whether the correction provides a higher number of good trials depends on the accuracy of the correction algorithm, which may not be sufficient particularly if the data contain a substantial number of drifts (i.e. slow voltage shifts, often caused by sweating) (Plöchl et al. 2012).

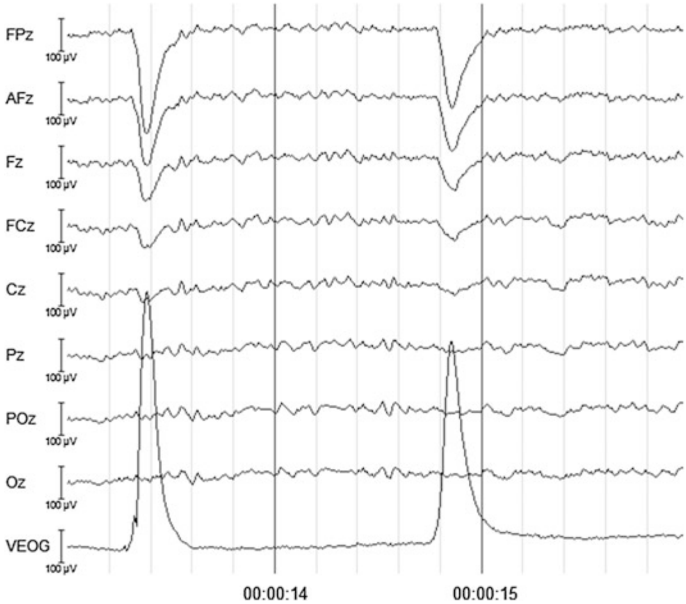

Fig. 6.5 Recordings from an electrode registering vertical eye-movements (VEOG) and EEG electrodes located at frontal, central and posterior cites. Two blinks can be seen slightly before the 14 and 15 s markers

Ground and online reference electrodes: In addition to the EOG electrodes which monitor eye movements and blinks, there are two more types of electrode which are not necessarily part of the cap but essential for the EEG registration system. The first of these is the ground electrode, which is a common reference point needed to be able to measure voltages. Different systems and different labs vary in their protocol as to where to place this electrode—for example, in some labs a position on the forehead is favoured, while others recommend placing it on the sternum. Additionally, some systems substitute this one ground electrode by two electrodes that are for example placed on a location in the cap. The exact location will not affect the overall readings. Secondly, in order to eliminate electrical noise from the amplifier's ground circuit, a reference electrode is necessary,[5] and the actual signal that is being measured is the voltage difference between this electrode and every other electrode on the head. The location of the reference can affect the scalp distribution of the overall signal, and if such differences should occur across locations, it will become harder to identify an overall effect. It is therefore wise to pay some attention to this aspect of the set-up. However, if it should not be possible to use the same reference at all locations, it is possible later on to re-reference the

[5]In some systems, e.g. Biosemi, data can be recorded reference-free, but has to be referenced afterwards in order to prevent unnecessary noise in the data.

signal to a different location (there are standard methods available in the software packages for EEG registration and analysis to conduct this re-referencing offline).

Sampling rate (SR) is usually expressed in hertz (Hz), or number of measurements per second. On many current amplifiers, the sampling rate can be set to a number of different rates, but this is not true of all set-ups. In general, an SR of 500 or 512 Hz is commonly used in language experiments. For analyses, data are often downsampled—i.e., averaged across a longer time window, but it is still useful to record at a high SR. The higher the original SR is, the more reliable the (averaged or downsampled) measurement, since the noise will be more reduced through the averaging of more data points. It is good to keep the SR the same across locations, but it is possible to downsample datasets with a higher SR in order to match them to datasets with a lower SR (the higher rate does not have to be an exact multiple of the lower one). Since the forms of analysis discussed in the next section do not depend on a high sampling rate, reducing the measured data intervals does not present a problem. In fact for some approaches downsampling provides a useful data reduction within the statistical analysis.

Preprocessing of the data after their acquisition is an essential step in the analysis of data acquired from several labs. This can be done with different kinds of software, from more programming-based (Fieldtrip in Matlab), to more user-friendly interfaces (EEGLAB in Matlab or Brain Vision Analyzer as stand-alone software). There are many different steps to be taken in preprocessing the data, and they are not always similar across labs. With multi-lab data, it is important to think about the reasons why some steps need to be taken. Since many of these preprocessing details to some extent depend on the analyzing software in your home lab, they will not be discussed in full detail here. Instead, we will give some examples of how we decided to process the data in the experiment discussed in this volume. Note that there are many technical details beyond the scope of this volume associated with each of these steps. We refer the reader to Luck (2014) for a full discussion.

- Downsampling: Datasets with higher sample rates were downsampled to 500 Hz to make all sampling rates equal.
- Re-referencing: We chose to re-reference our data to mastoid locations (or electrodes in the cap that were very close to the same position), because out of the possible options, these were the most similar across locations. In addition, since most language studies on sentence processing have made use of averaged mastoid or ear references, which are relatively similar to each other, this made our study comparable to previous research.
- Filtering: We filtered our data with a high-pass filter of 0.1 Hz, and a low-pass filter of 40 Hz. We chose to do this at 24 dB/octave, because this gives a slow slope of rejecting frequencies. Again, the choices depend on what is best for the data available and the effects you are interested in. Some researchers decide not to filter at all, but filtering the data removes some extremely high and low frequency alternations that can be regarded as noise.

- Ocular Correction: We decided to use the so-called Gratton-Coles algorithm for ocular correction. In order to keep this correction as similar as possible across locations we calculated bipolar EOG channels for those datasets that did not have a bipolar EOG measurement. Eye movement correction can occur before or after the segments of EEG time-locked to an event are extracted from the data. This depends on the algorithm which is used. With Gratton-Coles, using unsegmented, continuous data yields better results.
- Segmentation: With this step, segments of the data are created which include a baseline period, used to show that no differences existed between conditions before the event of interest. The time window following the event should include all time windows which are relevant. For example, the P600 may extend until 1400 ms after the information that evokes it, in monolinguals, so the epoch should be at least that long; to examine delayed effects, a longer epoch should be used when learner populations are included.
- Artifact rejection: We removed segments in which the activity violated certain criteria (e.g., a large increase in microvolts over 200 ms). There are also ways to correct artifacts instead of removing them, for example using an independent or principle components analysis (ICA or PCA) to detect certain regularities such as blinks, drifts or other artifacts. This procedure if often time-consuming and may not be worth the effort if you are interested in large components such as the P600. It is also more subjective than using clear criteria for artifact rejection.

6.6 Statistical Approaches and Interpretation of Results

There are generally three types of ***statistical procedures*** used for the analysis of EEG data. The first is the most traditional and commonly used statistical technique, namely univariate or repeated measures analysis of variance (ANOVA). The common procedure is to (1) create ERPs by averaging single trials of EEG (epochs) per individual per condition, (2) calculate the average/maximum amplitudes per individual within a specific latency window and (3) use these average/maximum amplitudes as a dependent variable in the ANOVA analysis to test hypotheses (Hoormann et al. 1998). This approach assumes that all the factors are factorial (e.g., that there are equal numbers of participants in each group) and is generally difficult to apply when continuous factors like age of onset/arrival are involved.

For this reason, mixed-effects or multiple regression approaches have been developed which can determine which factors are the best predictors of the learning outcome (e.g., Baayen 2008; Cunnings 2012; see Smith and Kutas 2015, for an application within EEG research). Mixed effects models in particular allow for the inclusion of participants with missing data. In addition, they can handle single trial analysis (as opposed to averaging across trials), non-linear time effects, time × person effects (random slope effects) and a within-subject covariate. A third, more recent, approach, generalized additive mixed-effects models (GAMs) is an

extension of the mixed effects regression approach which can deal with non-linear effects gracefully (Wood 2006; see Tremblay and Baayen 2010, for an application of GAMs to EEG data). For this reason it is an interesting method for examining delay in onset and other complex differences in waveform that are measured with ERPs.

A detailed treatment of the exact statistical approaches we took in the example experiment, which differed to some extent depending on the exact datasets being analyzed, is beyond the scope of this chapter. Instead the reader is referred to the citations above for more in-depth discussion. However, we want to encourage researchers to think carefully about the choice of a particular statistical method, since this can have important consequences for the outcome and conclusions of a study. Many ERP studies on bilingual processing are somewhat problematic in that they create discrete groups on the basis of continuous variables such as age of acquisition, rather than taking into account the full range of variability amongst learners (i.e., by using the numerical value itself). Not only are the decisions on where to set the cutoff age between the different groups invariably somewhat arbitrary, in order to be able to uncover the true shape of age of acquisition or attrition effects an approach that uses continuous predictors is also far superior. Most studies that have resorted to traditional group averaging might therefore not have been able to report the full story.

Lastly, we want to say a word about potential ***multilab data differences***—that is, effects in the data that are actually the outcome of them having been collected at different sites. In order to ensure that this is not the case, the researcher should ensure that there is no interaction between any effect of interest and the measuring site. A second problem for data analysis may be posed when the numbers and positions of the electrodes in the cap differ across labs. The best strategy here is to determine the locations of the electrodes in those caps that have the smallest overall number, and continue the analysis using only these locations, ignoring all other electrodes in those cases where there are more. However, sometimes all electrodes are not situated at exactly the same corresponding locations. In this case, a region of interest (ROI) based approach can be useful. In this approach, the signal from a number of electrodes which cover a certain area of the scalp where an effect is expected is combined. For example, the P600 is expected to be largest over parietal electrodes and to extend to central and occipital electrodes as well. Figure 6.6 shows a selection of electrodes that can therefore be used in a ROI that investigates P600 effects. Although some sensitivity to exact location may be lost when using ROIs, differences in scalp distribution are difficult to interpret across different groups, as they may reflect differences in the brain morphology (sulci and gyri) rather than a truly different location in the brain. Such a reduction of the data may also be necessary for some types of analyses, as GAM analyses of large numbers of electrodes or ROIs can take up very significant amounts of processing time.

When ***interpreting results*** it is important to keep in mind the real goal of the study. It is tempting to over-interpret minor effects of factors such as AoA or the L1 of the learner. If all bilinguals, regardless of AoA, look quite similar to the

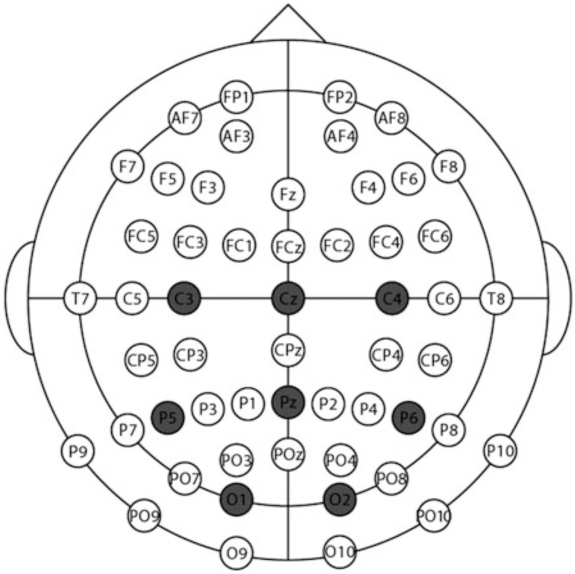

Fig. 6.6 An example of a typical region of interest (ROI) to measure the P600 effect

monolinguals (e.g. in that they display a P600), then this predictor plays a minor role in how a second language is processed, even if there is a statistically significant interaction. Quantitative differences, such as a delay in onset or a smaller amplitude of an ERP component, suggest processing limitations rather than truly different processing. The factors that cause qualitative changes are the most interesting: If an N400 appears instead of a P600, it indicates that a particular variable is associated with a truly different strategy (McLaughlin et al. 2010).

Suggestions for Further Reading

Handy, T.C. (ed.). 2005. *Event-related potentials: A methods handbook*. Cambridge: MIT press.
Kaan, E. 2007. Event-related potentials and language processing: A brief overview. *Language and Linguistics Compass* 1(6): 571–591.
Kutas, M., and C. Van Petten. 1994. Psycholinguistics electrified. Event-Related Brain Potential Investigations. In *Handbook of psycholinguistics*, ed. M.A. Gernsbacher, 83–143. San Diego: Academic Press.
Luck, S.J. 2014a. *An introduction to the event-related potential technique*, 2nd ed. Cambridge: MIT press.
Picton, T.W., S. Bentin, P. Berg, E. Donchin, S.A. Hillyard, R. Johnson, et al. 2000. Guidelines for using human event-related potentials to study cognition: Recording standards and publication criteria. *Psychophysiology* 37(02): 127–152.
Rugg, M.D., and M.G. Coles. 1995. *Electrophysiology of mind: Event-related brain potentials and cognition*. Oxford: Oxford University Press.

References

Baayen, H. 2008. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Baayen, R.H., Piepenbrock, R., and Gullikers, L. 1995. *The CELEX lexical database [CD-ROM]*, 682. Philadelphia: Linguistics Data Consortium, University of Pennsylvania.

Balota, D.A., M.J. Yap, M.J. Cortese, K.A. Hutchison, B. Kessler, B. Loftis, et al. 2007. The english lexicon project. *Behavior Research Methods* 39: 445–459.

Barber, H., and M. Carreiras. 2005. Grammatical gender and number agreement in Spanish: An ERP comparison. *Journal of Cognitive Neuroscience* 17(1): 137–153.

Besson, M., M. Kutas, and C. Van Petten. 1992. An event-related potential (ERP) analysis of semantic congruity and repetition effects in sentences. *Journal of Cognitive Neuroscience* 4(2): 132–149.

Brouwer, H., H. Fitz, and J. Hoeks. 2012. Getting real about semantic illusions: Rethinking the functional role of the P600 in language comprehension. *Brain Research* 1446: 127–143.

Clahsen, H., and C. Felser. 2006. Grammatical processing in language learners. *Applied Psycholinguistics* 27(01): 3–42.

Coulson, S., J.W. King, and M. Kutas. 1998. Expect the unexpected: Event-related brain response to morphosyntactic violations. *Language and Cognitive Processes* 13(1): 21–58.

Cunnings, I. 2012. An overview of mixed-effects statistical models for second language researchers. *Second Language Research* 28(3): 369–382.

Dimitrova, D.V. (2012). *Neural correlates of prosody and information structure*. PhD Thesis, University of Groningen.

Dimitrova, D.V., Redeker, G., and Hoeks, J.C. (2009). Did you say a BLUE banana? the prosody of contrast and abnormality in Bulgarian and Dutch. *Interspeech* 999–1002.

Dimitrova, D.V., L.A. Stowe, G. Redeker, and J.C. Hoeks. 2012. Less is not more: Neural responses to missing and superfluous accents in context. *Journal of Cognitive Neuroscience* 24 (12): 2400–2418.

Federmeier, K.D., and M. Kutas. 1999. A rose by any other name: long-term memory structure and sentence processing. *Journal of Memory and Language* 41: 469–495.

Foucart, A. (2008). *Grammatical gender processing in French as a first and second language*. PhD Dissertation, University of Edinburgh and University of Provence, France.

Foucart, A., and C. Frenck-Mestre. 2011. Grammatical gender processing in L2: Electrophysiological evidence of the effect of L1–L2 syntactic similarity. *Bilingualism: Language and Cognition* 14(03): 379–399.

Foucart, A., and C. Frenck-Mestre. 2012. Can late L2 learners acquire new grammatical features? Evidence from ERPs and eye-tracking. *Journal of Memory and Language* 66(1): 226–248.

Frenck-Mestre, C., A. Foucart, H. Carrasco, and J. Herschensohn. 2009. Processing of grammatical gender in French as a first and second language evidence from ERPs. *EuroSLA Yearbook* 9(1): 76–106.

Friederici, A.D. 1995. The time course of syntactic activation during language processing: A model based on neuropsychological and neurophysiological data. *Brain and Language* 50(3): 259–281.

Friederici, A.D., A. Hahne, and A. Mecklinger. 1996. Temporal structure of syntactic parsing: Early and late event-related brain potential effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 22(5): 1219.

Friederici, A.D., E. Pfeifer, and A. Hahne. 1993. Event-related brain potentials during natural speech processing: Effects of semantic, morphological and syntactic violations. *Cognitive Brain Research* 1(3): 183–192.

Gratton, G., and M.H. Coles. 1989. Generalization and evaluation of eye-movement correction procedures. *Journal of Psychophysiology* 3: 14–16.

Gunter, T., A. Friederici, and H. Schriefers. 2000. Syntactic gender and semantic expectancy: ERPs reveal early autonomy and late interaction. *Journal of Cognitive Neuroscience* 12(4): 556–568.

Hagoort, P., and C.M. Brown. 1999. Gender electrified: ERP evidence on the syntactic nature of gender processing. *Journal of Psycholinguistic Research* 28(6): 715–728.

Hagoort, P., L. Hald, M. Bastiaansen, and K.M. Petersson. 2004. Integration of word meaning and world knowledge in language comprehension. *Science* 304: 438–441.

Holcomb, P.J., and H.J. Neville. 1991. Natural speech processing: An analysis using event-related brain potentials. *Psychobiology* 19(4): 286–300.

Hoormann, J., M. Falkenstein, P. Schwarzenau, and J. Hohnsbein. 1998. Methods for the quantification and statistical testing of ERP differences across conditions. *Behavior Research Methods, Instruments, and Computers* 30(1): 103–109.

Institut für Deutsche Sprache. 2010. Deutsches Referenzkorpus/Archiv der Korpora geschriebener Gegenwartssprache 2010-I (Release vom 02.03.2010). Mannheim: Institut für Deutsche Sprache. http://www.ids-mannheim.de/kl/projekte/korpora/archiv.html.

Kaan, E., A. Harris, E. Gibson, and P. Holcomb. 2000. The P600 as an index of syntactic integration difficulty. *Language and Cognitive Processes* 15(2): 159–201.

Keuleers, E., K. Diependaele, and M. Brysbaert. 2010. Practice effects in large-scale visual word recognition studies: a lexical decision study on 14,000 Dutch mono- and disyllabic words and nonwords. *Frontiers in Psychology* 1: 1–15.

King, J., and M. Kutas. 1995. Who did what and when? Using word-and clause-level ERPs to monitor working memory usage in reading. *Journal of Cognitive Neuroscience* 7(3): 376–395.

Kluender, R., and M. Kutas. 1993. Bridging the gap: Evidence from ERPs on the processing of unbounded dependencies. *Journal of Cognitive Neuroscience* 5(2): 196–214.

Kuhl, P.K. 2010. Brain mechanisms in early language acquisition. *Neuron* 67(5): 713–727.

Kutas, M., and S.A. Hillyard. 1980. Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science* 207(4427): 203–205.

Kutas, M., T. Lindamood, and S. Hillyard. 1984. Word expectancy and event-related brain potentials during sentence processing. In *Preparatory states and processes*, ed. S. Kornblum, and J. Requin, 217–237. Hillsdale: Erlbaum.

Loerts, H. 2012. *Uncommon gender. Eyes and brains, native and second language learners, & grammatical gender*. PhD dissertation, University of Groningen.

Loerts, H., L.A. Stowe, and M.S. Schmid. 2013. Predictability speeds up the re-analysis process: An ERP investigation of gender agreement and cloze probability. *Journal of Neurolinguistics* 26(5): 561–580.

Luck, S.J. 2005. Ten simple rules for designing ERP experiments. In *Event-related potentials: A methods handbook*, ed. T.C. Handy. Cambridge: MIT Press.

Luck, S.J. 2014b. *An introduction to the event-related potential technique*, 2nd ed. Cambridge: MIT press.

Luck, S.J., and E.S. Kappenman (eds.). 2012. *The Oxford handbook of event-related potential components*. Oxford: Oxford University Press.

Malmivuo, J., and R. Plonsey. 1995. *Bioelectromagnetism: Principles and applications of bioelectric and biomagnetic fields*. New York: Oxford University Press.

McLaughlin, J., L. Osterhout, and A. Kim. 2004. Neural correlates of second-language word learning: Minimal instruction produces rapid change. *Nature Neuroscience* 7(7): 703–704.

McLaughlin, J., D. Tanner, I. Pitkänen, C. Frenck-Mestre, K. Inoue, G. Valentine, and L. Osterhout. 2010. Brain potentials reveal discrete stages of L2 grammatical learning. *Language Learning* 60(s2): 123–150.

Meulman, N., L.A. Stowe, S.A. Sprenger, M. Bresser, and M.S. Schmid. 2014. An ERP study on L2 syntax processing: When do learners fail? *Frontiers in Psychology* 5: 01072. doi:10.3389/fpsyg.2014.

Misra, M., T. Guo, S.C. Bobb, and J.F. Kroll. 2012. When bilinguals choose a single word to speak: Electrophysiological evidence for inhibition of the native language. *Journal of Memory and Language* 67(1): 224.

Molinaro, N., H.A. Barber, and M. Carreiras. 2011. Grammatical agreement processing in reading: ERP findings and future directions. *Cortex* 47(8): 908–930.

Molinaro, N., F. Vespignani, and R. Job. 2008. A deeper reanalysis of a superficial feature: An ERP study on agreement violations. *Brain Research* 1228: 161–176.

Ojima, S., H. Nakata, and R. Kakigi. 2005. An ERP study of second language learning after childhood: Effects of proficiency. *Journal of Cognitive Neuroscience* 17(8): 1212–1228.

Osterhout, L. 1997. On the brain response to syntactic anomalies: Manipulations of word position and word class reveal individual differences. *Brain and Language* 59: 494–522.

Osterhout, L., and P.J. Holcomb. 1992. Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language* 31(6): 785–806.

Osterhout, L., and J. Nicol. 1999. On the distinctiveness, independence, and time course of the brain responses to syntactic and semantic anomalies. *Language and Cognitive Processes* 14(3): 283–317.

Osterhout, L., P.J. Holcomb, and D.A. Swinney. 1994. Brain potentials elicited by garden-path sentences: Evidence of the application of verb information during parsing. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 20(4): 786.

Peltola, M.S., T. Kujala, J. Tuomainen, M. Ek, O. Aaltonen, and R. Näätänen. 2003. Native and foreign vowel discrimination as indexed by the mismatch negativity (MMN) response. *Neuroscience Letters* 352(1): 25–28.

Phillips, C., N. Kazanina, and S.H. Abada. 2005. ERP effects of the processing of syntactic long-distance dependencies. *Cognitive Brain Research* 22(3): 407–428.

Plöchl, M., J.P. Ossandón, and P. König. 2012. Combining EEG and eye tracking: identification, characterization, and correction of eye movement artifacts in electroencephalographic data. *Frontiers in Human Neuroscience* 6: 278.

Roehm, D., I. Bornkessel, H. Haider, and M. Schlesewsky. 2005. When case meets agreement: Event-related potential effects for morphology-based conflict resolution in human language comprehension. *NeuroReport* 16(8): 875–878.

Rossi, S., M. Gugler, A. Friederici, and A. Hahne. 2006. The impact of proficiency on syntactic second-language processing of German and Italian: Evidence from event-related potentials. *Journal of Cognitive Neuroscience* 18(12): 2030–2048.

Sabourin, L., and L. Stowe. 2008. Second language processing: When are first and second languages processed similarly? *Second Language Research* 24: 397–430.

Schneider, W., A. Eschman, and A. Zuccolotto. 2002a. *E-Prime user's guide*. Pittsburgh: Psychology Software Tools Inc.

Schneider, W., A. Eschman, and A. Zuccolotto. 2002b. *E-Prime reference guide*. Pittsburgh: Psychology Software Tools Inc.

Sharbrough, F., G.E. Chatrian, R.P. Lesser, H. Lüders, M. Nuwer, and T.W. Picton. 1991. American electroencephalographic society guidelines for standard electrode position nomenclature. *Journal of Clinical Neurophysiology* 8(2): 200–202.

Smith, N.J., and M. Kutas. 2015. Regression-based estimation of ERP waveforms: I. The rERP framework. *Psychophysiology* 52: 157.

Steinhauer, K., K. Alter, and A.D. Friederici. 1999. Brain potentials indicate immediate use of prosodic cues in natural speech processing. *Nature Neuroscience* 2: 191–196.

Steinhauer, K., E.J. White, and J.E. Drury. 2009. Temporal dynamics of late second language acquisition: Evidence from event-related brain potentials. *Second Language Research* 25(1): 13–41.

Swaab, T.Y., K. Ledoux, C.C. Camblin, and M. Boudewyn. 2012. Language-related ERP components. In *The Oxford handbook of event-related potential components*, ed. S.J. Luck, and E.S. Kappenman, 397–439. New York: Oxford University Press.

Tremblay, A., and R.H. Baayen. 2010. Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall. In *Perspectives on formulaic language: Acquisition and communication*, ed. D. Wood, 151–173. London: The Continuum International Publishing Group.

Van Berkum, J.J., C.M. Brown, and P. Hagoort. 1999. When does gender constrain parsing? Evidence from ERPs. *Journal of Psycholinguistic Research* 28(5): 555–566.

Van Hell, J.G., and N. Tokowicz. 2010. Event-related brain potentials and second language learning: Syntactic processing in late L2 learners at different L2 proficiency levels. *Second Language Research* 26(1): 43–74.

Van Petten, C., and M. Kutas. 1990. Interactions between sentence context and word frequency in event-related brain potentials. *Memory & Cognition* 18(4): 380–393.

Van Petten, C., and M. Kutas. 1991. Influences of semantic and syntactic context on open-and closed-class words. *Memory & Cognition* 19(1): 95–112.

Weber, K., and A. Lavric. 2008. Syntactic anomaly elicits a lexico-semantic (N400) ERP effect in the second language but not the first. *Psychophysiology* 45(6): 920–925.

Weber-Fox, C., and H. Neville. 1996. Maturational constraints on functional specializations for language processing: ERP and behavioral evidence in bilingual speakers. *Journal of Cognitive Neuroscience* 8(3): 231–256.

Weber-Fox, C., and H.J. Neville. 2001. Sensitive periods differentiate processing of open-and closed-class words: An ERP study of bilinguals. *Journal of Speech, Language, and Hearing Research* 44(6): 1338–1353.

Wood, S. 2006. *Generalized additive models: An introduction with R*. Boca Raton: Chapman & Hall/CRC Press.